UNIVERSIDADE DE LISBOA

Faculdade de Ciências| Faculdade de Letras| Faculdade de Medicina|Faculdade de Psicologia

Inteligência Artificial e Consciência Fenoménica. Quão perto estamos de máquinas conscientes?

José António Rodrigues do Carmo

Dissertação orientada pela Profª. Doutora Adriana Graça e pelo Prof. Doutor Helder Coelho, especialmente elaborada para a obtenção do grau de Mestre em Ciências Cognitivas.

2017

UNIVERSIDADE DE LISBOA

Faculdade de Ciências| Faculdade de Letras| Faculdade de Medicina|Faculdade de Psicologia

Inteligência Artificial e Consciência Fenoménica. Quão perto estamos de máquinas conscientes?

José António Rodrigues do Carmo

Dissertação orientada pela Profª. Doutora Adriana Graça e pelo Prof. Doutor Helder Coelho, especialmente elaborada para a obtenção do grau de Mestre em Ciências Cognitivas.

2017

“A Consciência de existir, a raiz
Do ilimitado, omnímodo mistério
Que tem tronco de Ser, folhas de vida
Flores de sentimento e sofrimento
E frutos do pensar, podres depressa.

A Consciência de existir, tormento
Primeiro e último do raciocínio
Que, porém, filho dela, a não atinge.
A Consciência de existir me esmaga
Com todo o seu mistério e a sua força
De compreendida incompreensão profunda,
Irreparavelmente circunscrita.”

Fernando Pessoa, *A Consciência de existir, a raiz*

Agradecimentos

Este trabalho está escrito, por opção, segundo a norma ortográfica antiga. Não por qualquer discordância activista, mas pelo facto de a nova ortografia me deixar sempre com uma vaga e desagradável impressão de erro. Consequência provável de ter aprendido a língua num tempo em que cada erro correspondia a uma reguada.

Confessada esta minha perturbação pós-traumática, há muitas pessoas a quem devo agradecer.

Aos docentes do Mestrado em Ciência Cognitiva, das várias Faculdades da Universidade de Lisboa, evidentemente. Tive o privilégio de beneficiar dos seus profundos conhecimentos e do seu entusiamo e agradeço-lhes, não só os ensinamentos que me transmitiram durante a formação académica, mas também, e sobretudo, a disponibilidade e a atitude de incentivo que perpassou todos os actos pedagógicos.

Todos me ajudaram a tentar rasgar horizontes mais amplos relativamente ao tema e simultaneamente a aprofundar a imensa humildade científica que se sente quando se estudam os mistérios da mente humana, dando substância à sábia ideia de que quanto mais sabemos, mais sabemos que pouco sabemos.

A verdade é que todos, sem excepção, me permitiram que os incomodasse para pedir conselhos e ajuda, e responderam sempre com presteza, precisão, simpatia e honestidade.

Aos colegas, pelas conversas, dúvidas, apoios e amizade, neste breve tempo em que coincidimos nos caminhos da vida e do conhecimento.

Agradeço a excepcional e exigente orientação propiciada pela Professora Doutora Adriana Silva Graça e pelo Professor Doutor Helder Coelho, reputados académicos, cujos saberes alcançam uma latitude de campos nada comum, num tempo em que a crescente especialização tende a fazer afunilar as perspectivas nos extremismos analíticos. Os seus incisivos, cirúrgicos e exigentes comentários, bem como a constante indicação de referências, foram essenciais para manter a abordagem do assunto no nível de análise adequado, e se melhor não fiz, foi exclusivamente por incapacidade própria para aproveitar tudo o que me deram.

À minha filha, o propósito maior da minha vida.

À Margarida, cuja memória estará sempre presente e que ao partir na última viagem me deixou, como derradeiro gesto de amor, as dúvidas existenciais que me trouxeram aqui!

Resumo

A consciência é um fenómeno mal compreendido de sistemas biológicos com uma certa complexidade, e há quem sustente a possibilidade de poder vir a ser uma propriedade de determinados organismos artificiais. Alguns autores acreditam que pode emergir do substracto físico, ser uma ilusão ou um mero epifenómeno pelo que, teoricamente, nada obsta a que possa ser instanciada, uma vez compreendidos os mecanismos que a fazem surgir. Para outros, entre os quais me incluo, a instanciação da consciência em organismos artificiais não é possível, uns porque a consideram irredutível ao físico, outros porque a situam em planos transcendentais.

Neste trabalho, para além de considerações gerais destinadas a situar o assunto, procuro abordar particularmente o conceito de consciência fenoménica, o chamado "problema duro", o problema de saber como é que certas actividades neuronais aparecem internamente como experiência subjectiva, como *qualia*, à luz de diferentes teorias oriundas de vários campos da ciência. Para além das referências às principais teorias metafísicas, discuto com algum pormenor as mais relevantes teorias específicas da consciência, analiso modelos e implementações propostas pela Inteligência Artificial (IA) e pela Consciência Artificial (CA)[1], discuto até que ponto se avançou, ou não, na simulação e na instanciação da consciência em organismos artificiais, e quais as principais objecções à sua instanciação e caracterização.

Por fim são extraídas algumas conclusões que tentam responder à questão suscitada no título, partindo da ideia de que a consciência fenoménica não é processamento de informação, de uma intuição *a priori*[2] de que não é uma propriedade emergente do substracto físico, e que talvez só seja possível em determinados organismos biológicos.

Palavras-chave: Consciência; experiência subjectiva; *qualia*; instanciação; simulação; intuição.

1 Consciência artificial, "*machine consciousness"* ou até "senciência digital". Embora o segundo termo pareça estar a ganhar terreno, neste trabalho prefiro usar a designação "Consciência Artificial".

[2] Para (BonJour, 1998) a intuição é a base da justificação *a priori*.

Abstract

Consciousness is a poorly understood phenomenon of both more and less complex biological systems, and there are those who defend the possibility that it may also become a property of certain artificial organisms. Some authors believe that consciousness can arise from the physical substrate, be an illusion or a mere epiphenomenon, so theoretically nothing prevents it from being implemented once the mechanisms that make it emerge are understood. For others, myself included, the implementation of consciousness in artificial organisms is not possible, either because it is considered to be irreducible to the physical, or because it is situated in transcendental order.

In this work, in addition to general considerations aimed at placing the subject, I particularly seek to address the concept of phenomenic consciousness, the so-called "hard problem," the problem of how, in the light of different theories from various fields of science, certain neural activities appear internally as subjective experience, or *qualia*. In addition to references to the main metaphysical theories, I discuss in some detail the most relevant specific theories of consciousness, analyze models and implementations proposed by Artificial Intelligence and Artificial Consciousness, and discuss to what extent progress has been made in the simulation and instantiation of consciousness in artificial organisms, and what are the main objections to its instantiation and characterization.

Finally, some conclusions are drawn in an attempt to answer the question raised in the title, starting from the idea that phenomenic consciousness is not a processing of information, and from an a priori intuition that it is not an emergent property of the physical substrate, and that perhaps it is only possible in certain biological organisms.

Keywords: Consciousness; subjective experience; *qualia*; implementation; simulation; intuition.

Índice de conteúdos

Índice de figuras

Lista de abreviaturas, siglas e acrónimos

ACH- (Arquitectura Cognitiva de Haikonen)

ADN- (Ácido desoxirribonucleico)

BRAIN- (Brain Research through Advancing Innovative Neurotechnologie)

CA- (Consciência Artificial)

CERA-CRANIUM- (Conscious and Emotional Reasoning Architecture-Cognitive Robotics Architecture Neurologically Inspired)

CNC- (Correlatos neuronais da consciência)

CODAM- (Corollary Discharge of Attention Movement)

EEG- (Electroencefalograma)

ETG- (Espaço de Trabalho Global)

fMRI- Functional Magnetic Resonance Imaging

HAL- (Heuristically programmed Algorithmic)

HBP- (Human Brain Project)

IA- (Inteligência Artificial)

IDA- (Intelligent Distributed Agent)

MEG- (Magnetoencefalografia)

MRI- (Magnetic Resonance Imaging)

MT- (Máquina de Turing)

NDE- (Near death experience)

Orch-OR- (Orchestrated Objective Reduction)

RNA- (Rede neuronal artificial)

TETG- (Teoria do Espaço de Trabalho Global)

TII- (Teoria da Integração da Informação)

TMS- (Transcranial magnetic stimulation)

TT- (Teste de Turing)

TTT- (Teste de Turing Total)

"O facto fundamental do universo é haver consciência dele."
[Fernando Pessoa]

1 Introdução

Em 1968, no filme "*2001-Odisseia no Espaço*", de Kubrik, sobre um guião de Arthur C. Clarke, HAL 9000[3] teve "consciência" de que ia ser desligado "atemorizou-se" e decidiu eliminar os humanos. À época acreditava-se que estavam para breve máquinas capazes, não só de imitar todas as capacidades cognitivas humanas, mas também de ter sensações, sentimentos e consciência de si e dos seus processos. A caminho dos finais do séc. XX, para a maioria dos investigadores a consciência era ainda algo ao alcance da ciência, particularmente da computação e da neurociência, mas o fenómeno revelava-se elusivo e complexo, com resultados muito aquém das expectativas.

Passados 50 anos, muitos dos que mantém a esperança quanto à factibilidade de máquinas conscientes, transitaram para um certo pessimismo cautelar quanto às suas eventuais consequências, temendo uma ameaça existencial à espécie humana. Especula-se até com a "singularidade", a possibilidade de máquinas inteligentes e com intencionalidade própria, se imporem como espécie dominante, numa espiral endógena de cada vez mais rápidas melhorias, introduzidas por sucessivas gerações de máquinas (Vinge, 1993), (Hawking, S., 2017).

É realmente possível a consciência em máquinas? Para começar a esboçar uma resposta, é necessário sublinhar que não há consenso sobre o que é a consciência, quais os mecanismos que a geram e porque e como surge no mundo biológico. Apesar disso a maioria dos seres humanos não duvida que tem consciência, que tem experiências subjectivas conscientes e que elas são centrais na sua vida.

A tese funcionalista, predominante no actual paradigma científico, é a de que a consciência pode ser um fenómeno que emerge da interacção funcional de várias partes do cérebro, ideia de algum modo descrita pela poética metáfora de Damásio (1999) comparando a consciência com um maestro gerado pela própria orquestra quando começa a tocar. Recorrendo a extraordinárias imagens funcionais do cérebro, a neurociência vem procurando isolar os chamados correlatos neuronais da consciência (CNC), mas a descrição física completa de um sistema não é suficiente para explicar a experiência subjectiva, e não se sabendo como o cérebro "produz" a consciência, também não se sabe como ou se outros tipos de sistemas a poderiam "produzir".

Alguns entendem que a consciência, seja o que for, é apenas uma propriedade dos sistemas biológicos (Searle J. , 2016), de máquinas especiais que são os cérebros. Outros acreditam que qualquer sistema pode vir a ser consciente, desde que tenha a estrutura certa. Chalmers, (2017), por exemplo, entende que a consciência emerge (nomológica mas não metafísica ou conceptualmente) da

[3] O *Heuristically programmed ALgorithmic* (HAL) 9000, era o computador da nave e foi conceptualizado por Marvin Minsk.

organização física e funcional, e pensa que sistemas artificiais poderão ser conscientes, pelo menos como questão de necessidade nomológica, mesmo que a consciência não seja físicamente explicável. Assume também a possibilidade de leis psicofísicas que liguem o processamento da informação à consciência, e o físico ao mental.

Esboçar uma resposta à questão deste trabalho implica pois um pequeno mergulho na Filosofia, na IA, na Psicologia, na Neurociência e talvez na Física, para tentar delimitar o conceito de consciência e indagar das respostas já avançadas pelos investigadores, acredite-se ou não na possibilidade de as máquinas poderem a vir a desenvolver uma consciência fenoménica.

Sim, porque no estado actual do conhecimento trata-se basicamente de uma questão de crença. Para os que não acreditam, como é o meu caso, a resposta é intuitiva e *a priori*: não são possíveis máquinas conscientes e por isso, do ponto de vista pragmático, a questão poderia nem sequer fazer sentido. Todavia em ciência as questões fazem sempre sentido, por vezes são até a única coisa que faz sentido, e importa seguir o seu fio de Ariadne para tentar escapar do labirinto solipsista. Para os que acreditam no poder da ciência, sentimento legítimo porque sustentado nos extraordinários avanços que ela tem propiciado, esta, leve o tempo que levar, acabará por deslindar o novelo e esclarecer os mecanismos físicos que levam à consciência.

2 Estrutura do trabalho

O trabalho começa com uma revisão das principais questões filosóficas sobre a consciência, e procura isolar o problema duro. Mostrarei que este problema está relacionado com os *qualia*, os quais são tratados com algum detalhe.

Seguidamente abordam-se perspectivas que tentam colocar o problema da consciência no âmbito do estudo da engenharia, particularmente da IA e da Consciência Artificial (CA). Podem as máquinas ter *qualia*? Há quem sugira que sistemas sem *qualia* não são verdadeiramente conscientes, pelo que se discute também o que poderão vir a ser os *qualia* artificiais.

Uma vez que se coloca o problema de avaliar até que ponto um organismo é consciente, apresentam-se e discutem-se brevemente alguns testes e provas correntemente usados e/ou propostos.

No capítulo 6 descrevem-se algumas das mais importantes teorias específicas da consciência e as aplicações relevantes, com o objectivo de perceber em que medida abordam a consciência fenoménica e se aproximam da solução do problema duro

No capítulo 7 discute-se a informação referida nos capítulos anteriores, enfatizando as principais objecções à possibilidade de CA e, no capítulo 8 são extraídas conclusões.

3 Consciência

3.1 Mente, consciência e experiência subjectiva

Qualquer que seja a definição adoptada (e há muitas), a mente não tem substância física, nem existe no espaço. Engloba percepções, observações, imaginação, pensamentos, vontade, emoções, raciocínios e todos os processos mentais, conscientes ou inconscientes, que lhes subjazem, além do sentido do "eu" que torna a mente consciente de si mesma.

Está evidentemente correlacionada com o cérebro, mas a possibilidade de descrever processos mentais sem ter de descrever os subjacentes neuronais, sugere desde logo que os processos mentais podem ser executados por diferentes mecanismos, da mesma maneira que um *software* pode correr em diferentes *hardwares*. Esta analogia foi uma das hipótese fundacionais da CA, no que toca à possibilidade de organismos artificiais poderem vir a ter uma mente consciente.

A consciência parece exigir a experiência subjectiva[4]. Ao escrever estas linhas posso ver e sentir os músculos a trabalhar. Mas quando penso, não vejo nem sinto os disparos dos potenciais de acção nos meus neurónios. Só tenho a experiência subjectiva desses processos, a qual não ocupa lugar, não é ponderável nem constituída por partículas. Os modernos instrumentos mostram representações simbólicas sobre um cérebro que se ativa de certas maneiras, mas não mostram a imagem, a dor, o cheiro, a cor, o som, o medo, que o organismo está a experienciar subjectivamente. Os meus pensamentos e percepções sobre objectos no mundo, não são objectos físicos dentro da minha cabeça e o que experiencio não é o mundo físico lá fora. Este, como sabemos, é descrito por teorias assentes em objectos, reais ou imaginários, que procuram mostrar que as mudanças que percebemos são provocadas por interacções entre esses objectos[5], mas o que realmente experiencio é apenas uma representação do mundo físico, eventualmente gerada pelo meu sistema sensorial. Uma alteração das condições químicas do cérebro pode levar, perante o mesmo mundo físico, à experiência de um mundo fenoménico inteiramente diferente. Sob o efeito de psicotrópicos posso experienciar diferentes formas, cores, etc. O que se altera não é o mundo físico, mas a representação que dele faz o meu cérebro, pelo que o fluxo da consciência acontece, todo ele, no mundo fenoménico. E todavia o meu cérebro é um objecto físico, descrito por teorias físicas, constituído por partículas, com massa, dimensões, etc. De uma forma intrigante, a experiência fenoménica parece estar associada a este dispositivo físico, embora não se conheça o mecanismo dessa associação.

[4] Segundo Chalmers, (1995), quando pensamos e percepcionamos há uma agitação de processamento de informação mas também um aspecto subjectivo, que é experiência. Quando vemos, experienciamos sensações visuais: A qualidade da vermelhidão, a experiência da escuridão e da luz, etc. Depois, há sensações corporais, imagens mentais, a qualidade sentida da emoção, etc. O que une todos estes estados é que há algo que é como estar neles. São estados experienciais.

[5] Por exemplo, a alteração da cor de um composto líquido será motivada por interacções moleculares.

A consciência permite-me estar imerso no mundo sem qualquer esforço de processamento. Mas em vez de internalizar o mundo percepcionado, externalizo as percepções sensoriais para que pareçam o mundo. A externalização é perfeita porque, apesar de toda a informação ser uma representação produzida no cérebro, o mundo parece-me efectivamente lá fora e não dentro da minha cabeça. Além disso há uma clara sensação de continuidade, um fluxo de experiência que me garante que um objecto é o mesmo quando o experiencio de diferentes posições ou em diferentes tempos.

A aparência subjectiva é interna, só está disponível para mim, mesmo que seja a de um mundo físico lá fora, eventualmente partilhado por outros. A consciência é sobre objectos e condições, reais ou imaginárias, não sobre os processos neuronais, químicos ou físicos que lhes subjazem. Mas não resulta automaticamente da actividade neuronal. Eu tenho várias actividades neuronais que não são acompanhadas de qualquer experiência subjectiva, pelo que também uma eventual máquina com um cérebro electrónico não terá necessariamente experiências subjectivas.

Enfim, posso dizer que estou consciente quando experiencio dor, prazer, medo, cores, sons, quando observo o mundo e os estados do próprio corpo, quando tenho um discurso interno e elaborado sobre ele. Estou consciente até quando sonho, apesar de adormecido. Em resumo, estou consciente quando experiencio, independentemente de qualquer intencionalidade.

3.2 Como definir a consciência?

Não existe uma definição consensual de consciência e Searle (2016) considerou uma boa definição do senso comum chamar consciência a todos aqueles estados em que se tem experiências subjectivas, senciência ou vigília, e que começam todos os dias quando acordamos de um sono sem sonhos e se sucedem até voltarmos a cair no sono, morrermos ou perdermos a consciência de qualquer outra forma. Esta definição contém óbvios elementos de circularidade e procurando uma maior objectividade, Block, (1995 e 2002) propôs os conceitos de:

- Consciência de acesso (consciência A), funcional, ligada à atenção, encarando um estado consciente como uma representação apta a ser livremente usada no raciocínio e controle directo de outras acções.

- Consciência de monitorização (consciência M), que inclui os processos de percepção interna ou introspecção, isto é, tem a ver com a capacidade de um organismo gerar um modelo interno de si mesmo;

-Autoconsciência (Consciência S) que é a capacidade de autorreconhecimento e de raciocínio sobre o que se reconhece, enfim a consciência de um "eu";

-Consciência fenoménica (consciência P) que refere o conjunto de experiências subjectivas que um organismo tem pelo facto de ser consciente como, por exemplo uma cor, um som, uma dor, um

cheiro, tempo, espaço, etc. A consciência P é-nos familiar mas difícil de definir. Para um organismo, há algo que é ser como é (Nagel T. , 1974)[6]. Block entende que é distinta de qualquer propriedade cognitiva, intencional ou funcional e Aleksander (2009) também concorda que a consciência pode não ter a ver com inteligência ou capacidades cognitivas[7].

Boltuc (2009) propôs uma divisão ainda mais simples, entre a consciência H (*hard*), puramente fenoménica, e a consciência F (funcional): ao pôr um dedo numa chama, experiencio dor (consciência H) e retiro a mão (reacção funcional). A dor ensina-me também a não repetir o gesto e isto é algo que se pode fazer numa máquina, mas sem a real experiência da dor, bastando definir a dor artificial como um processo que realiza todos os aspectos funcionais da dor. A consciência H é a experiência subjectiva, a "verdadeira consciência". Os artefactos que implementam as reacções funcionais não são necessariamente conscientes. Harnad & Scherzer (2007) concordam que ser consciente é apenas experienciar.

Tanto Boltuc como Harnad & Scherzer entendem que a consciência fenoménica não tem qualquer função e todas as funções que lhe são atribuídas se devem aos processos neuronais que subjazem à experiência consciente, e não à experiência em si, pelo que esta pode ser um mero epifenómeno, como o ruído de um motor.

Face à consciência fenoménica há quem conclua que o fisicalismo falha (Chalmers, 2003). Outros defendem-no metafisicamente mas argumentam que há um hiato explicativo entre o cérebro e a consciência fenoménica (Levine, 2014).

Seja o que for, a consciência fenoménica aparenta esconder-se no meio dos processos biológicos embora estes possam teoricamente passar sem ela, como sugere o argumento dos "*zombies*"[8] (Chalmers, 1996).

3.3 O problema duro da consciência e o hiato explicativo

O que faz com que determinados organismos sejam conscientes? Como é possível que uma consciência possa emergir de um objecto material e, reciprocamente, como é possível que influa no movimento de objectos materiais? E que vantagens confere a consciência aos organismos que a possuem?

O homem pensa sobre estas questões há milhares de anos e muitos entendem que a consciência envolve um problema mente-corpo, de difícil ou mesmo impossível resolução.

[6] Excluído deste trabalho está o conceito de "*awareness*", traduzível como "vigília". Muitas vezes, quando conduzimos estamos vigilantes relativamente a tudo o que se passa no trânsito, mas não temos consciência disso e nem nos damos conta de como chegámos a casa. Ademais, os casos de *blindsight* mostram claramente que existe vigília sem consciência. Também Damásio (1999) propõe que não existe qualquer implicação entre vigília e consciência dando como exemplo os sonhos.

[7] Embora admita a possibilidade de uma inteligência do tipo humano só poder ser instanciada com a presença dos processos que permitem a experiência subjectiva.

[8] Ente hipotético externamente indistinguível de um ser humano normal mas sem *qualia.* Um *zombie* não experienciará dor, mas reagirá como se sim.

A mente implica a experiência de estar consciente, mas como explicá-la nos termos das leis da Física que conhecemos? O cérebro biológico é um complexo sistema de células neuronais, glia, sinapses, etc., onde se dão os processos físicos sem os quais parece não haver experiência subjectiva[9]. Estes processos podem ser monitorizados com instrumentos e há correlações fortes entre processos neuronais e experiências subjectivas reportadas pelos organismos conscientes.

Porque se acredita que, tal como acontece no caso biológico, organismos artificiais conscientes poderão ser mais versáteis e adaptáveis a ambientes complexos, a ciência e a tecnologia ambicionam criar máquinas conscientes. Tratam o assunto como um projecto de engenharia, embora até ao momento falte aos engenheiros uma especificação operativa da experiência de estar consciente. O desconhecimento dos eventuais processos físicos na base do aparecimento de uma mente consciente e aparentemente imaterial, a partir de um cérebro, todo ele a funcionar segundo estritos processos físicos, faz-nos desembocar naquilo que Chalmers (1996) designou como o "problema duro" da consciência, o problema de explicar como a experiência fenoménica pode emergir do substracto físico[10], o problema de saber como é que certas actividades neuronais aparecem internamente como experiência subjectiva.

Sublinhe-se que nem todos encontram aqui um problema. Investigadores como Dennett (2017) entendem que não existe qualquer problema duro, que um cérebro é apenas uma máquina, e a consciência uma ilusão, um "Teatro Cartesiano"[11].

3.4 Teorias metafísicas da consciência

Uma teoria satistafória da consciência tem de enquadrar todas as particularidades conhecidas do fenómeno nos seres humanos, explicar o caso dos outros animais, dar conta dos casos clínicos que a comprometem e, sobretudo, vencer o hiato explicativo, resolvendo o problema duro. A Filosofia tem tentado fazê-lo e a questão central é se o fenoménico e o físico são duas diferentes realidades ou substâncias, ou se se podem reduzir uma à outra.

3.4.1 Idealismo

O idealismo não aceita a realidade metafísica do mundo físico e afirma a prevalência do mental sobre o físico. Berkeley (1988) dizia que as ideias eram a única realidade e Husserl (1960) sugeria

[9] Certos casos de NDE (*Near Death Experience)* parecem colocar em cheque esta afirmação (Alexander E, 2012), mas pelo menos a sua reportabilidade não escapa à exigência de processos físicos.

[10] Chalmers referiu também os problemas "fáceis": explicar como somos capazes de processar informação, focar a atenção, reportar estados mentais, etc. Estes problemas, não sendo realmente fáceis, permitem uma aproximação teórica mais amigável.

[11] Para Dennett, ao eliminar o dualismo o que resta do modelo de Descartes é um teatro dentro do cérebro, onde um homúnculo observa os dados sensoriais que aí desembocam, toma decisões e dá ordens.

que deveríamos suspender a crença num mundo físico e enfatizar a descrição da experiência fenoménica, no sentido de nela alicerçar a ciência.

Sendo verdade que a maioria das pessoas, incluindo os cientistas, considera que o mundo físico está fora da consciência e é independente dela, considerando que ninguém tem ou teve a experiência de algo sem estar consciente desse algo, não parece todavia descabido conjecturar que qualquer coisa que seja experienciada está dentro da consciência e não fora.

O idealismo é uma teoria com consistente lógica interna, mas impossível de refutar segundo o actual paradigma científico, pelo que não desenvolveu até hoje um quadro de experimentação testável e no qual se possa trabalhar em termos de eventual implementação artificial.

3.4.2 Dualismo

Para Platão existia um mundo material e um mundo das ideias e Descartes (1637) sugeria também que a mente e o corpo eram substâncias diferentes, este regulado por leis físicas e aquela uma entidade imaterial em interacção com o cérebro.

Isto importa à CA. As máquinas conscientes deverão também ter uma mente que contenha a imaterial experiência subjectiva, um conteúdo mental que igualmente parecerá imaterial ao organismo artificial.

O dualismo cartesiano conduz de imediato ao problema da interacção entre a mente e o corpo. Ou a maquinaria física é apenas um aparelho que sintoniza a consciência, ou tem de haver algo, uma entidade, um homúnculo, um conversor que, a jusante e a montante da mente imaterial, faça as conversões matéria / não matéria, necessárias para que os sinais físicos sensoriais se convertam em experiências subjectivas e para que as intenções, pensamentos, desejos etc., accionem o corpo e os seus processos físicos. Este conversor, que para Descartes se situava na glândula pineal[12], tem de ser capaz de criar energia e/ou matéria a partir do nada.

A crítica ao dualismo cartesiano não aceita a premissa da imaterialidade de mente e entende que esta nos parece imaterial apenas porque não percepcionamos a maquinaria que lhe subjaz, já que o cérebro não reconhece os seus próprios padrões neuronais de forma a determinar que este ou aquele padrão representa esta ou aquela entidade ou conceito.

Como a ciência não lida bem com conceitos que escapem às leis físicas conhecidas, a ideia de mente imaterial não pode ser comprovada, pelo que se considera que o dualismo cartesiano escorrega para fora dos limites do método científico. Contudo, e numa peculiar ironia científica, o estranho mundo da física quântica sugere, numa possível interpretação, que se está constantemente a criar

[12] Não há qualquer prova de que a glândula pineal seja um canal de comunicação entre o físico e o mental. Sabe-se apenas que está fortemente correlacionada com os ritmos circadianos.

energia e matéria literalmente a partir do nada, apenas por flutuações quânticas no espaço-tempo (Sokolov et al., 2010).

3.4.3 Epifenomenalismo

Sendo embora intuitiva a diferença fundamental entre corpo e mente, entre processos físico, químicos, neuronais e experiência subjectiva, o epifenomenalismo[13] rejeita a existência de diferentes substâncias e considera o mundo fenoménico como sendo um epifenómeno do mundo físico, sem qualquer influência causal sobre ele. Um fenómeno mental seria, *mutatis mutandis*, como a sombra de um objecto: só existe se o objecto está presente e há luz[14].

O principal problema do epifenomenalismo é que torna difícil falar de estados fenoménicos, já que as descrições da consciência geradas pelo cérebro não estão causalmente ligadas aos estados fenoménicos, e por isso não podem ser sobre eles.

Do ponto de vista da CA o epifenomenalismo parece de pouca utilidade porque não havendo redutibilidade às propriedades físicas não é possível nessa base projectar um mecanismo que leve à emergência.

3.4.4 Teorias da Identidade

Estas teorias resolvem o problema da emergência afirmando que a mente é apenas a acção dos neurónios e suas ligações, e negando que existam fenómenos com propriedades mentais.

A identidade tipo-tipo propõe que um determinado tipo de actividade neuronal (por exemplo a estimulação de certas fibras) corresponde sempre a um certo tipo de actividade mental (por exemplo, a dor). A identidade *token-token* propõe que um certo tipo de actividade mental pode ser gerada por diferentes tipos de actividades neuronais, da mesma maneira que num computador uma dada imagem pode ser obtida por diferentes algoritmos. Esta variante interessa particularmente aos engenheiros pela possibilidade de múltipla realização, segundo a qual as propriedades mentais poderão ser realizáveis em diversos substractos físicos (Fodor, 1974).

3.4.5 Fisicalismo e funcionalismo

Para o fisicalismo só existe o mundo material, descrito por leis físicas, a mente é um complexo de funções de processamento de informação, realizadas pelo computador "cérebro" e a consciência tem

[13] Também designado "dualismo de propriedades".

[14] O dualismo de propriedades fundamentais considera as propriedades mentais conscientes como constituintes básicos da realidade, a par de propriedades físicas fundamentais, como a carga eletromagnética, por exemplo. Estas propriedades mentais podem interagir com outras mas ontologicamente a sua existência não depende nem deriva delas (Chalmers D. , 1996). Para o Dualismo de propriedades emergentes as propriedades conscientes emergem de organizações complexas de componentes físicos e não são *a priori* previsíveis nem explicáveis a partir da natureza estritamente física (Hasker, 1999). O Dualismo de propriedades monista neutral entende que as propriedades mentais conscientes e as propriedades físicas derivam de um ainda mais fundamental nível da realidade que não é uma coisa nem outra (Strawson, 1994), embora seja discutível que este monismo neutral possa ser classificado como um dualismo de propriedades.

algo a ver com processamento de informação, funções e estruturas da matéria física. Não há outra causalidade que não a física. Dennett (2017) descreve o cérebro como uma máquina feita de biliões de máquinas (os neurónios), e defende que tudo é produto do acaso, de forças aleatórias geradas pela matéria e pela evolução, natural e cultural. Os fenómenos mentais são ilusões. Quando olho para a minha caneta, pode parecer-me que estou a ver umas formas e umas cores no meu subjectivo campo visual, mas trata-se de uma ilusão, já que a única realidade é um processo físico que está a ocorrer no meu córtex visual, e que não consigo descrever.

Os estados mentais[15] são identificados pelo que fazem e não por aquilo de que são feitos. O funcionalismo entendido como *fisicalismo causal,* sustenta que as funções são propriedades que actuam sobre a realidade através de mecanismos causais que realizam fisicamente a propriedade abstracta. Por exemplo, a propriedade «ser um afiador de facas" é realizada por vários mecanismos físicos possíveis, de plástico, metal, pedra, etc. O fisicalista defende pois que uma boa explicação funcional equivale a uma explicação baseada em leis causais e o que importa é a estrutura do afiador de facas, não aquilo de que é feito. Analogamente todo o sistema físico com as propriedades de um sistema neuronal que gera a suporta a consciência, suportá-la-á também.

A matéria está associada com estados fenoménicos quando se organiza num cérebro, mas não quando se organiza numa cadeira, por exemplo, e por isso um estado ou processo é consciente em função do papel funcional que desempenha num sistema devidamente organizado. A visão funcionalista é especialmente relevante para a CA, na variante do **funcionalismo da máquina** (Putnam, 1967), segundo a qual os estados mentais são apenas estados funcionais de um Máquina de Turing, aberta também à possibilidade da múltipla realização. Uma acção intencional, por exemplo beber uma "mini", resulta de dois estados mentais: o desejo de a beber e a crença de que há uma no frigorífico. Segundo a interpretação fisicalista os dois estados mentais são estados informacionais que actuam causalmente sobre o corpo, porque os seus realizadores (redes neuronais ou circuitos semicondutores), são os que agem realmente sobre os músculos. Como disse Putnam, não é preciso descer ao nível quântico para explicar porque é que um parafuso quadrado não entra numa porca redonda.

3.4.6 Pampsiquismo

Para o pampsiquismo os constituintes básicos do universo tem propriedades mentais e a consciência é uma propriedade primordial da estrutura da realidade, sendo que as propriedades físicas são partes menores das mentais. Nagel (2012) considera que o pampsiquismo é uma espécie de

[15] Compostos ou realizados por estados físicos, numa relação inter-níveis em que propriedades ou factos de um nível são realizados por complexas interacções entre itens do nível inferior, como por exemplo, a relação entre o atómico e o químico

dualismo de propriedades e liga a sua ressurgência[16] ao falhanço do emergentismo, uma vez que entende que não há verdadeiras propriedades emergentes dos sistemas complexos[17].

É uma teoria antiga, bastante transversal, e nas suas diferentes versões colhe defensores em vários campos, da ciência e não só. Podem considerar-se no âmbito do pampsiquismo certas teorias específicas de que falarei adiante, como a Teoria de Integração da Informação (TII) (Tononi, 2008), a *Orchestrated Objective Reduction* (*Orch-OR*) (Hameroff & Penrose, 2014)[18], e a Teoria do Holofluxo.

Uma das principais objecções ao pampsiquismo é que não há evidência de que as entidades físicas tenham propriedades mentais, situação caricaturada por (Searle, 1997), ao considerar absurdo que um termostato possa ter consciência.

[16] O pampsiquismo é uma teoria antiga, subscrita por filósofos como Parménides, Platão, Espinoza, Leibniz, William James, etc. Há elementos pampsiquistas nas filosofias orientais como o Taoísmo, Budismo, etc.
[17] São, todas elas, derivadas das relações entre eles e algo e das propriedades dos seus componentes e seus efeitos uns sobre os outros, quando combinados (Nagel, T, 1979).
[18] O mental é ontologicamente fundamental no Universo. (Penrose R. , 1997).

4 O mundo fenoménico

4.1 *Qualia* e experiência subjectiva

Eu não percepciono os meus neurónios a disparar quando transportam informação sobre o mundo. A sua acção aparece-me directamente como *qualia*, como experiência subjectiva de qualidades e objectos do mundo e do corpo, cores, sabores, humidade, sons, dor, prazer, tempo, espaço, etc. Estes *qualia* são inefáveis, intrínsecos, privados, estruturados e irredutíveis[19].

Os *qualia* estão directamente relacionados com o problema duro. Como é que os padrões de actividade neuronal são experienciados internamente como *qualia* e não como aquilo que são? Ou como é que são experienciados, pura e simplesmente? Desde logo, os *qualia* não parecem ser propriedades do mundo real, mas sim gerados pelo sistema sensorial[20]. O fenómeno da sinestesia comprova que certos indivíduos, mercê de disfunções do sistema sensorial, cruzam *qualia* de diferentes modalidades sensoriais (Neckar & Bob, 2016) e qualquer pessoa pode notar diferenças no que vê, ouve ou cheira, por exemplo, quando sob o efeito de alucinogénios.

E todavia os *qualia* têm uma conexão com o mundo. Nas mesmas condições internas e externas um certo espectro visível evocará um certo *output* sensorial que aparecerá ao organismo consistentemente como um certo *quale*, uma certa sensação de cor. Designar uma cor por "amarelo" é apenas descrever o *quale* que experiencio, se o vir ou imaginar. Para Jackson (1982), os *qualia* não são físicos nem podem ser criados por meios físicos, porque toda a informação relacionada com um percepto pode ser descrita sem eles. No seu argumento[21] apresenta a cientista Mary que conhece todo o processo fisiológico de ver cores. Mary nunca viu outras cores que não o branco e o preto, pelo que se for confrontada com outra cor, experienciará o seu *quale*, embora isso nada acrescente ao seu prévio conhecimento sobre a visão das cores, ou seja, os *qualia* não acrescentarão informação física à que Mary já tem, logo não são físicos. Contudo se me disserem que ao colocar um dedo no fogo me irei queimar e retirar a mão, etc., também saberei tudo o que irá acontecer, mas colocar lá o dedo, embora tecnicamente nada acrescente à informação, provocará um *quale* (dor), o que inequivocamente acrescenta conhecimento que não estava na informação descritiva. Ou seja, informação sem *qualia* é apenas descrição, mas os *qualia* não requerem descrição, não se podem resumir a símbolos ou valores comparáveis. Experienciam-se. Vermelho é aquilo que experiencio como vermelho, independentemente do nome que lhe dou.

Explicar a consciência fenoménica é, no fundo, explicar como se geram os *qualia*.

[19] Inefáveis, no sentido em que só podem ser apreendidos pela experiência directa; intrínsecos, porque se trata de propriedades não relacionais; privados porque não se podem comparar com os *qualia* de outros agentes; irredutíveis, porque a experiência de um *quale* contém tudo o que há para saber sobre esse *quale*; estruturados, porque parecem ter uma estrutura que permite comparações. Os *qualia* correspondentes a diferentes modalidades sensoriais são diferentes, mas mesmo assim podem comparar-se a suavidade de uma cor com a suavidade de um toque.
[20] Como se comprova com a excitação artificial de sensores usada, por exemplo, em implantes cocleares (Ling & Spencer, 2017).
[21] Conhecido por "*knowledge argument"*

4.2 A privacidade dos *qualia*

Os *qualia* são subjectivos e apenas referenciáveis a outros por descrição indirecta. Experiencio o meu *quale* de amarelo, não o de outras pessoas, e apenas a semelhança biológica me autoriza a especular que pode ser parecido com o meu, embora jamais possa ter a certeza. Na verdade pessoas daltónicas podem chamar vermelho ao que eu chamo vermelho, e não estão provavelmente a ter o mesmo *quale*, ou seja, a descrição simbólica nada nos diz sobre a experiência subjectiva.

Quanto à imagiologia, os instrumentos com que estudamos o cérebro detectam processos físicos e produzem símbolos e gráficos. Mas os *qualia* não são esses símbolos e, até à data, nenhum *qualia* ou estado mental foi detectado directamente. Mesmo que pudesse ligar o meu cérebro ao de outra pessoa, provavelmente experienciaria os meus *qualia*, sem garantias de que seriam iguais aos dela.

As diferentes modalidades sensoriais dão origem a *qualia* típicos, apesar de o subjacente (neurónios a disparar) ser semelhante, pelo que a via neuronal seguida pode ser relevante. Uma certa via sensorial estará associada a um certo *quale*. Uma cor não é experienciada como um som, excepto em casos de sinestesia, ou uso de substâncias psicotrópicas.

Há contudo algumas propriedades comuns a várias modalidades sensoriais, tais como formas, ritmo, etc. Haikonen (2012) designa-as por propriedades amodais, experienciadas como *qualia* amodais, cuja representação neuronal parece estar sintonizada com o mundo. Por exemplo, o ritmo de uma peça musical é uma característica do fenómeno, mas também da actividade neuronal que o processa e que parece ser partilhado por vários organismos da mesma maneira. Se ouço um ritmo os meus neurónios disparam a esse ritmo e o meu pé bate no chão ao mesmo ritmo. O *quale* do ritmo manteve-se através de todo o sistema. Isto significa que podemos ter uma ideia de como certos *qualia* amodais serão eventualmente experienciados por organismos artificiais, se os tiverem. É certo que Nagel (1974) argumentou que não podíamos experienciar como era ser morcego, mas para certos *qualia* talvez essa impossibilidade não seja absoluta, porque as propriedades amodais parecem não depender apenas dos mecanismos que as transportam.

5 Inteligência Artificial e Consciência Artificial

5.1 Consciência humana *versus* consciência artificial

Encarar o cérebro como uma rede neuronal, implica que, por definição, não pode correr programas. Todavia o cérebro pode executar qualquer programa de computador com a simples ajuda de um papel e um lápis, e este paradoxo mostra que há diferenças relevantes entre uma Rede Neuronal Artificial (RNA) e uma rede neuronal biológica, como parece ser, em parte, o cérebro.

Segundo um paradigma da IA a mente baseia-se em computações simbólicas (Minsky, 2006). Por outro lado a abordagem conexionista sugere que a cognição e a consciência assentam em processamento sub-simbólico (os *qualia* são representações sub-simbólicas, dor é dor), que não pode ser executado por computadores digitais. O cérebro parece operar principalmente ao nível sub-simbólico mas com mecanismos que lhe permitem passar para representações simbólicas, algo que não acontece nas RNA. Haikonen (2012) sugere que, no cérebro, as representações simbólicas e sub-simbólicas são padrões neuronais semelhantes, mas os usados como símbolos têm um significado associado.

Nos humanos a consciência fenoménica parece assentar em *qualia*, mas muitas das funções que normalmente se atribuem à consciência são realizáveis por funções cognitivas, as quais podem ser artificialmente implementadas sem a necessidade de *qualia*. Apesar disso considera-se que um método de processamento de informação que contenha *qualia* poderia melhorar a execução das funções cognitivas e é essa uma ideia que subjaz à investigação em CA. Sendo assim, os sistemas artificiais conscientes deverão ser, tal como os humanos, capazes de interacção directa com o mundo, com a ajuda da experiência memorizada. Deverão experienciar directamente as qualidades do mundo e do próprio corpo, tal como nós, embora com *qualia* provavelmente diferentes.

Ao examinar o estado da arte na CA confrontamo-nos com trabalhos baseados em muitas perspectivas filosóficas, com diferentes objectivos e usando diversos métodos computacionais. Há no entanto temas recorrentes que permitem organizar esses estudos em categorias assentes naquilo que os diversos investigadores entendem ser fundamental para a consciência. Os modelos computacionais referidos neste trabalho têm em comum o facto de gerarem alguma expectativa quanto à instanciação da CA. Não é um critério de inclusão perfeito, já que por vezes os avanços surgem como aspectos colaterais de outras pesquisas, mas delimita de algum modo o campo de exploração.

Nos últimos anos vêm-se testando algum teorias específicas da consciência com modelos computacionais, especulando-se se isso pode levar a organismos artificiais fenomenicamente conscientes. Procura-se inspiração nos processos dos seres humanos e outros mamíferos superiores, tendo em vista criar sistemas artificiais que exibam capacidades e funcionalidades análogas e que, no limite, as tenham mesmo. E, nesse processo, procura-se também compreender melhor como se produz

a consciência no cérebro humano. O que se faz é, pragmaticamente, partir do que se teoriza e observa sobre o funcionamento da consciência humana para a formulação de modelos computacionais os quais, por sua vez, podem ajudar no estudo daquela.

Segundo Gamez (2008) o campo de pesquisa da CA pode organizar-se em 4 grandes áreas:

- Comportamento associado à consciência

Apesar de termos muitos comportamentos não conscientes, certos comportamentos complexos exigem-nos consciência. E é geralmente com base no comportamento que julgamos da presença de consciência noutros organismos.

- Capacidades cognitivas associadas à consciência

É consensual que existem correlações entre consciência e capacidades cognitivas, e apesar de não estar demonstrada qualquer ligação necessária entre cognição e estados fenoménicos[22] Aleksander & Dunmall (2003) e Aleksander (2005) sugeriram que uma eventual aproximação cognitiva à CA se baseasse em 5 axiomas que consideram minimamente necessários para que possa existir consciência: representação[23], imaginação[24], atenção[25], planeamento[26] e emoção[27].

A modelação destas capacidades tem sido feita de várias maneiras.

- Arquitecturas que se acredita causarem ou estarem correlacionadas com a consciência humana

A pesquisa nesta área surge do desejo de modelar e testar teorias neuronais e cognitivas da consciência, como a Teoria do Espaço de Trabalho Global (TETG) ou a Teoria da Integração da Informação (TII). Acredita-se que uma certa disposição funcional dos componentes físicos é necessária para simular a consciência e, no limite, instanciá-la.

As suas linhas de contacto com as áreas anteriores e com a pesquisa da consciência fenoménica são difusas mas investigadores como Haikonen (2012), por exemplo, sugerem que a implementação de certas arquitecturas pode mesmo vir a instanciar estados fenoménicos.

[22] Efectivamente atribuímos intuitivamente estes últimos a numerosos organismos biológicos sem que, todavia, pareçam satisfazer as pré-condições axiomáticas.
[23] O sistema tem estados perceptuais que representam coisas no mundo e a sua localização.
[24] O sistema pode recordar partes do mundo, ou criar sensações que são como partes do mundo.
[25] O sistema é capaz de escolher que partes do mundo representa ou imagina.
[26] O sistema usa sequências de estados para planear acções
[27] Tem estados afectivos que avaliam acções planeadas e determinam a sua prossecução

- Organismos com consciência fenoménica.

Aqui trata-se de tentar criar organismos verdadeiramente conscientes e não apenas que simulem ser conscientes pela replicação de comportamentos, capacidades ou arquitecturas. Parte-se da hipótese (ou esperança) de que a implementação em organismos artificiais, de comportamentos, estados cognitivos e arquitecturas internas, pode eventualmente vir a produzir experiências fenoménicas reais.

Por outro lado também há quem entenda ser possível vir a instanciar a consciência sem nada do anterior. Tononi & Koch (2015), alegam que até um díodo pode ter um grau de consciência, na medida em que integra informação. Se estas alegações estiverem correctas a presença de estados fenoménicos num organismo artificial poderá não depender apenas das funções de alto nível que executa, nem dos comportamentos, nem das arquitecturas, mas de algo mais que por ora se desconhece.

5.2 *Qualia* na máquina

Os organismos conscientes têm *qualia* pelo que, para Block (1978), uma teoria que não contemple os *qualia*, não serve para explicar a consciência. Qualquer organismo artificial consciente terá, tal como nós, de experienciar a realidade de forma directa, como propriedades aparentes do mundo e não através de representações simbólicas indirectas, ou como os sinais físicos dos mecanismos que lhes subjazem.

Assim como o projecto fenomenológico (Husserl, 1960) visava a descrição da consciência humana, a fenomenologia sintética, conceito introduzido por Jordan (1988), ambiciona caracterizar os estados fenoménicos possuídos ou modelados por organismos artificiais (Chrisley, 2009).

Coloca-se sempre o problema da 1ª pessoa. Nos humanos a experiência interior de outros humanos infere-se essencialmente por empatia e observação de comportamentos. Não será o caso dos organismos artificiais nos quais, tal como acontece com os animais, terá de partir sobretudo de enfoques baseados na 3ª pessoa, que consistem em examinar o comportamento, incluindo formas de reporte ou comunicação precisa (Seth et al., 2005), e também na inspecção da arquitectura e mecanismos internos do organismo. Esta abordagem aos *qualia* artificiais é usada de diversas maneiras em modelos e implementações computacionais, (Arrabales et al., 2010).

Para descrever a fenomenologia de um qualquer sistema, é condição *sine qua non* identificar aqueles que são capazes de estados fenoménicos. É para isso que servem os testes referidos em 5.3. Identificado um organismo com esta capacidade é necessário descrever os estados fenoménicos, quando e se ocorrerem. No nosso caso usamos a linguagem natural, mas não é óbvia a sua adequação a outros organismos. Por isso Gamez (2006) propôs um método para parcelar os estados internos de um sistema numa série de representações estruturadas ligadas a específicos estímulos ambientais.

Estas estruturas descrevem os estados, momento a momento, sem recurso a conceitos em linguagem humana.

Um outro método, de Stening et al. (2005), produz representações gráficas dos estados internos de um organismo, imprimindo a informação sensoriomotora armazenada nos seus conceitos.

Chrisley (2009) propõe a representação gráfica, baseada em expectativas, do "conteúdo não conceptual". Um exemplo é o *robot* SEER-3 (Chrisley & Parthermore, 2007) que produz (fig. 1) representações dinâmicas do conteúdo da experiência visual. Representações do mesmo tipo podem ser feitas para conteúdos conceptuais, afectivos, temporais, etc.

Figura 1- Representação gráfica baseada em expectativas: Robot SEER-3. Uma cor numa dada localização indica a expectativa do agente de receber essa cor como *input*, quando olha nessa direcção. A ausência de expectativas é indicada pela cor cinzenta e as regiões onde foi detectada mudança, estão a vermelho. (Chrisley & Parthermore, 2007)

Para Haikonen (2012) os computadores só poderão experienciar *qualia* se os métodos de processamento de sinal e a aquisição padrão de dados digitais conservarem as características amodais das coisas. Ora os sinais de *input* são convertidos numa representação simbólica (código binário) que exige informação adicional sobre o significado dos dígitos. Mas como as representações simbólicas não transportam *qualia*, Haikonen (2009), sugere um mecanismo de produção de *qualia* com símbolos cujo significado se alicerça na situação do organismo no mundo que o rodeia.

Um outro trabalho relevante é a análise das correlações computacionais dos *qualia* artificiais (Chella & Gaglio, 2009): um processo activo integra os fluxos de informação interna e externa para reconstruir uma visão subjectiva da cena percebida pela máquina que, segundo ao autores, é um q*uale* artificial do sistema. Sublinham todavia que usam o conceito de forma algo metafórica, designando por *quale* artificial um estado que se postula ser uma experiência fenoménica do *robot*, sem tecer considerações sobre se o *robot* experiencia verdadeiramente.

Figura 2- Imagem reconstruída a partir do registo da actividade neuronal. Adaptado de (Nihisimoto et al, 2011)

Outro exemplo é o trabalho de Kamitani & Tong (2005) que usam padrões de intensidade em *voxels*[28] para prever estados fenoménicos e seus conteúdos.

Por seu lado Nishimoto et al. (2011) conseguiram associar imagens captadas por uma câmara com os padrões de activação obtidos por fMRI (*Functional Magnetic Resonance Imaging*) de um sujeito a olhar para o mesmo objecto, e depois, na inversa, traduzir padrões de activação em imagens (fig. 2).

Os *qualia* também se podem caracterizar do ponto de vista da TII, de que falarei adiante. Balduzzi & Tononi (2009) propõem uma representação matemática para caracterizar as relações existentes no "espaço dos *qualia*" do ponto de vista da informação que estes integram, tendo em conta que o cérebro humano tem muitos estados não fenomenicamente conscientes e pode acontecer o mesmo com os organismos artificiais que venham a ser consideradas capazes de ter estados fenoménicos (fig. 3). A teoria identifica as sub-redes com maior "consciência", e qualifica de fenoménicos os seus estados internos.

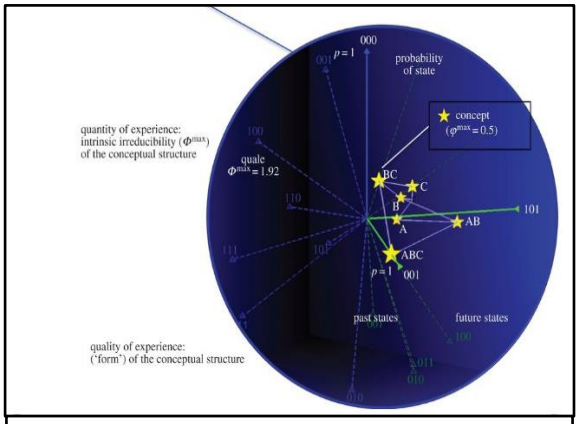

Figura 3- Espaço dos *qualia*, TII: Quantidade e qualidade da consciência num sistema simples. O *quale* é a forma da estrutura conceptual (Tononi & Koch, 2015)

5.3 Como saber se um organismo é consciente?

Nos humanos, para além de métodos que combinam a observação na 3ª pessoa[29], com o reporte do próprio, tendemos a supor, por analogia, que os outros seres humanos são conscientes da mesma maneira que nós. Com os animais, especialmente os não mamíferos, a questão não nos parece tão óbvia e com os organismos artificiais tudo se torna mais complicado (Prinz, 2003) porque a consciência fenoménica não pode ser detectada com instrumentos científicos, estabelecendo-se a sua presença por reportes verbais na 1ª pessoa, ou através da interpretação de comportamentos, como acontece com os animais. Mas há vários problemas com a inferência de consciência a partir do comportamento, mesmo em organismos biológicos. Para começar não é fiável, especialmente quando há lesões cerebrais[30]. Acontece também que qualquer pessoa pode estar a experienciar sem que haja comportamentos externos, como por exemplo num sonho ou em certos casos de *locked in*. Ou, na inversa, pode exibir comportamentos sem experienciar fenomenicamente, como fazem os actores, certos animais e máquinas especialmente desenhadas.

[28] "Voxel" é um pixel tridimensional. Cada pixel numa imagem de MRI corresponde a um voxel 3D no cérebro.
[29] Nos humanos os estados de consciência, no sentido de vigília, são por vezes caracterizados em termos das características definidoras anteriormente referidas ou através de comportamentos mensuráveis como o Glasgow e outras escalas de coma (Posner et al., 2007),
[30] Por exemplo, pacientes com anosognosia dizem ser capazes de usar um membro paralisado e confabulam para explicar a sua imobilidade.

Consequentemente têm-se procurado critérios objectivos, que confirmem ou infirmem a presença de consciência em organismos artificiais, para lá das inferências comportamentais, como a inspecção dos mecanismos internos do sistema.

Em humanos, um promissor teste, sugerido por Casali, et al. (2014), é independente do processamento sensorial e do comportamento, combinando EEG (electroencefalograma) e TMS (*Transcranial magnetic stimulation*) para quantificar o nível de consciência, tendo por base o grande reportório de padrões de activação e a integração (um sistema neuronal comportando-se como uma unidade singular).

O clássico Teste de Turing (TT) para a IA tem sido usado como referência para testar projectos mas a sua validade é questionável (Clowes & Seth, 2008), havendo até quem afirme que nada tem a ver com consciência (Haikonen P. , 2007a).

Por outro lado um organismo com uma maquinaria cognitiva que inclua percepção directa e não simbólica, conteúdo mental, introspecção, atenção, memória e retrospecção, respostas e reportes, pode ter condições necessárias para ser funcionalmente consciente mas não há forma de provar que isso seja suficiente para experienciar *qualia.*

Que dizer dos testes habitualmente sugeridos?

5.3.1 Teste de Turing Total

Já lá vai o tempo em que se pensava que uma conversa em linguagem natural ou um jogo de estratégia eram apenas atributos humanos e que, em si, eram prova suficiente de consciência. Hoje temos organismos artificiais que conversam entre eles, que podem passar o clássico TT, que ganham aos campeões humanos de Go e Xadrez, que vencem o *Jeopardy*, que se reconhecem num espelho, etc., e nenhum deles mostra evidência de ser fenomenicamente consciente ou ter algo que se possa chamar um "eu"[31]. Harnad (1992) propôs um Teste de Turing Total (TTT), segundo o qual um organismo consciente tem de ter uma *performance* exactamente igual à de um ser humano em qualquer teste cognitivo, perante qualquer juiz, durante todo o tempo.

Todavia este teste, que à primeira vista parece invulnerável, contém uma falácia lógica: é verdade que qualquer organismo cognitivamente semelhante a nós passará o teste, mas também é logicamente possível que o superam organismos cognitivamente semelhantes a nós embora não conscientes. Não basta argumentar que o TTT poderá ser tão exaustivo que nenhum outro artefacto que não seja cognitivamente semelhante aos humanos o passará. É preciso demonstrar a impossibilidade de que ele seja superado por organismos não conscientes e cognitivamente diferentes. E como se demonstra isso?

[31] Sublinhe-se que eu mesmo não tenho maneira alguma de provar objectivamente a outrem que não sou um *zombie*, já que posso fingir todos os comportamentos.

No fundo, o TTT é, tal como o clássico TT, um teste comportamental, pelo que não permite testar directamente a presença de *qualia.*

5.3.2 Teste de compreensão de uma imagem

Pareceria, por exemplo, que um bom teste para a consciência implicaria, por exemplo, pedir a um organismo artificial para descrever uma cena complexa do mundo real, de um modo que diferenciasse os seus aspectos chave. Nós fazemos isso muito bem. Perante uma simples foto somos capazes de descrever o que está a acontecer, por muito bizarra e distorcida que a imagem seja e elaborar sobre ela, sobre os objectos, sobre as relações causais entre eles e entre eles e o mundo, fazendo até juízos morais[32] e tudo isso de modo que seja considerado por humanos como razoável e provável.

Segundo a TII a informação consciente é integrada e unificada. Se a integração desaparece, como acontece quando dormimos, estamos anestesiados, etc., a consciência também. Um organismo precisa de um grande reportório de informação activamente conectada para estar consciente. Koch & Tononi (2011) propuseram o uso desta condição, na compreensão de imagens, para testar a consciência em máquinas: uma imagem mostra um ecrã de computador com um teclado em frente; outra mostra o mesmo ecrã com uma flor em frente. Segundo os autores, um organismo não consciente não encontrará nada de errado na 2ª imagem, por não evocar informação sobre a relação entre os itens da figura. Não compreenderá o que capta. Os humanos, pelo seu lado, reconhecem logo a incongruência, recorrendo à integração de informação memorizada.

Todavia este é, no fundo, também um TT, já que o organismo tem apenas de exibir capacidades cognitivas semelhantes às nossas, mostrando que se comporta como nós. Nada neste teste mostra que experiencia *qualia* e está realmente consciente.

5.3.3 ConsScale

Arrabales et al. (2010) sugerem que a consciência pode ser gradual, começando num grau mínimo e acabando numa possível superconsciência em máquinas. Para avaliar a consciência (funcional) em máquinas, propuseram uma escala biologicamente inspirada que começa no nível -1 (organismo sem corpo, por exemplo um aminoácido de uma proteína) e acaba no nível 11, que tem vários fluxos de consciência, sendo que o nível 10 é o nível humano adulto, e o nível 6 o de um bebé de 1 ano.

A ConsScale é útil e avalia satisfatoriamente o grau da consciência funcional que acreditamos existir em certos organismos, mas quanto a organismos artificiais apenas nos esclarece sobre as suas capacidades cognitivas. Não aborda a consciência fenoménica e por isso também não prova que o organismo seja fenomenicamente consciente.

[32] Por exemplo: um ladrão está assaltar a pobre senhora, indefesa, para lhe levar a mala, e ninguém faz nada, isto não se admite, etc.

5.3.4 Escala de Probabilidade

Para determinar se um sistema pode ter estados fenoménicos Gamez (2005) propôs uma escala que ordena arquitecturas e implementações de acordo com a probabilidade de serem capazes de estados fenoménicos. Por exemplo, a escala atribui à população da China, a funcionar como um cérebro humano, interligada por rádios e satélites (experiência conceptual sugerida por Block (1978)), uma pontuação de 786 em 812. Tem contudo o cuidado de referir que esta escala é apenas a formalização de uma intuição. Ora parece-me intuitivamente implausível que se possa atribuir ao organismo "China", qualquer capacidade de experienciar subjectivamente o que quer que seja, apesar de uma tão elevada pontuação na escala de Gamez.

Em suma, embora estes e outros critérios sejam interessantes, parcialmente úteis e operacionalizáveis, nenhum deles é geralmente aceite como tendo suficiente objectividade para confirmar a presença ou ausência de CA. São, todos eles, alvo de refutações substanciais, para além das que aqui sublinhei, e que podem ser consultadas em (Seth, 2009).

6 Teorias específicas da consciência e aplicações relevantes

São muitas as teorias que procuram dar respostas a características particulares da consciência. A *Stanford Encyclopedia of Philosophy* Van Gulick (2016) agrupa-as em grandes categorias, que incluem algumas das que considero mais relevantes, tendo em vista as implementações no domínio da CA.

6.1 Modelos baseados nas teorias de ordem superior.

Estas teorias defendem que um estado mental é fenomenicamente consciente se é (ou pode ser) objecto de uma representação [33] de ordem superior. Esquematicamente, o estado mental M é consciente se estiver acompanhado por um estado de ordem superior, cujo conteúdo é de que o organismo está em M (Rosenthal, 2005)[34]. Ou seja, a actividade mental consciente usa um nível mais alto de representação do que a actividade mental não consciente. Assim, desejar conscientemente um chocolate, implica estar nos estados mentais de desejar o chocolate e de ter tal desejo.

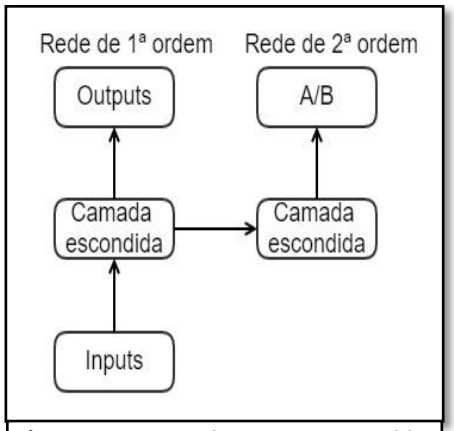

Figura 4- Uma arquitectura metacognitiva elementar. As caixas representam camadas de neurónios e as setas representam ligações entre todos os neurónios, cujos pesos mudam durante o treino (adaptado de (Cleeremans et al, 2007)).

As arquitecturas típicas têm uma rede de 1ª ordem para a tarefa primária, e uma rede de 2ª ordem, tida como funcionalmente consciente, que observa os estados da anterior e neles se baseia para tomar decisões (tarefa secundária). A fig. 4 esquematiza uma rede cuja tarefa primária é classificar *inputs*. A rede de 2ª ordem recebe apenas entradas da camada escondida da rede de 1ª ordem, e classifica de alta (A) ou baixa (B) a correcção de cada *output* da rede de 1ª ordem, face ao *input* que recebe. A medida em que a rede de 2ª ordem classifica correctamente o trabalho da rede de 1ª ordem, é interpretada como a confirmação de que o sistema está "consciente" da representação interna da rede de 1ª ordem. Cleeremans et al. (2007) afirmam que a experiência consciente ocorre se e só se um sistema de processamento de informação aprende acerca da sua própria representação do mundo. Neste modelo a rede de 2ª ordem aprende em simultâneo com a rede de 1ª ordem e vai mudando com ela.

[33] Estas representações podem ser percepções, pensamentos, ou crenças (Carruthers, 2000).
[34] A teoria tem variantes: a dos pensamentos de ordem superior (Rosenthal, 2005), e a das percepções de ordem superior (Lycan W. , 1996). As teorias reflexivas, por alguns consideradas também uma variante (Gennaro, 2012), localizam a consciência S directamente no conteúdo fenoménico de 1ª ordem do estado consciente e não num meta-estado (Van Gulick, 2004)

Num recente estudo comportamental de Persaud et al. (2007), um indivíduo com *blindsight*[35] fez apostas sobre a correcção de suas "suposições" acerca da presença ou ausência de estímulos visuais, naquilo que se pretendia ser uma medida objectiva da consciência do sujeito, relativa ao seu próprio desempenho. Pasquali et al. (2010) usaram a arquitectura metacognitiva ilustrada na fig. 5 para simular tais apostas pós-decisão.

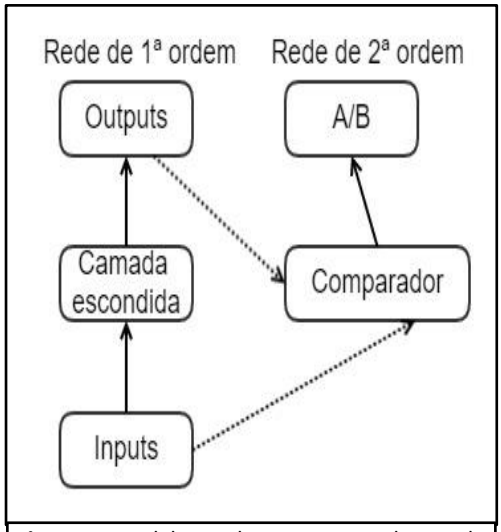

Figura 5- Modelo usado para o caso de estudo de *blindsight.* As setas ponteadas representam conexões não adaptativas (adaptado de (Pasquali et al., 2010))

Os nós da camada oculta na rede de 2ª ordem (comparador) recebem conexões de pares correspondentes de nós de entrada e saída da rede de 1ª ordem. As duas redes foram treinadas simultaneamente, a primeira para localizar correctamente os estímulos de entrada e a segunda para produzir boas decisões sobre decisões. O desempenho do modelo foi razoavelmente semelhante ao caso humano. Embora os autores não afirmem que o modelo instancia a consciência, sugerem que as representações de ordem superior formadas neste e noutros modelos metacognitivos capturam os mecanismos essenciais das teorias de pensamentos de ordem superior. Contudo Seth (2008) entende que nada têm a ver com medidas directas da consciência, e é bastante duvidoso que esta teoria possa abordar o problema duro e explicar os *qualia*

- Num trabalho sobre o CICEROBOT, um *robot* móvel usado como guia de museu, Chella & Macaluso (2006) alegam que a origem da autoconsciência está nas representações (de ordem superior) das percepções do mundo (de baixo nível) e ligam a comparação que o *robot* faz, entre as percepções reais e esperadas, com a presença de estados fenoménicos. Chella et al. (2008) utilizaram três níveis de representação: uma "área subconceptual" que lida com o processamento de baixo nível de dados sensoriais, uma "área conceptual" que organiza os dados sensoriais de nível inferior em categorias conceptuais, e uma "área linguística" de nível superior. A área conceptual, pré-simbólica, que alicerça os símbolos[36] usados na área linguística, é uma rede semântica de símbolos e suas relações com as percepções e acções do *robot*. É nela que ocorrem as inferências lógicas preditivas, estabelecendo-se expectativas para eventos subsequentes. Quando o *robot* se move, envia uma cópia dos comandos

[35] Fenómeno no qual indivíduos com danos no córtex visual primário respondem correctamente a estímulos visuais que não podem ver conscientemente.
[36] Este "alicerçar dos símbolos" refere-se ao conceito de "*symbol grounding*", que tem a ver com o significado atribuído a símbolos que teoricamente não passam de formas, cores, etc. Para Fodor (1985) o significado dos símbolos vem da sua adequada ligação ao mundo.

motores (*corollary discharge*[37]) para o simulador, que calcula expectativas acerca da próxima localização e da imagem da câmara.

- Num outro projecto Kitamura et al. (2000) estudaram um controlador robótico com uma arquitectura de cinco níveis. O nível mais baixo é um módulo reactivo e o mais alto (mais consciente) é um módulo simbólico, baseado em regras, que determina as estratégias globais de movimento. A abordagem foi bem sucedida no controlo dos comportamentos de dois *robots* em situação de busca e captura. Os autores sugerem que um "eu" consciente pode emergir nos níveis mais altos do *robot* quando as tarefas se tornam mais intencionais, *i.e*, quando as acções automáticas e reflexivas dos níveis inferiores são bloqueadas, os níveis mais altos entram em acção e simulam a direcção consciente da tarefa.

- Mais recentemente, a arquitectura cognitiva de quatro níveis CERA-CRANIUM [38] foi implementada em *robots* simulados e físicos Pioneer 3DX (Arrabales et al., 2010). Embora originalmente inspirados pela TETG, de que falarei a seguir, os autores sugerem que os *qualia* e a experiência consciente podem emergir no nível cognitivo mais elevado porque, em vez de aceder directamente aos dados sensoriais, o organismo processa indirectamente as percepções e trata-as como se estivessem espacialmente localizadas no mundo exterior.

Até ao momento nenhuma aplicação concreta produziu resultados encorajadores. Afinal de contas até os computadores que temos em casa incluem sistemas de monitorização dos seus estados internos o que, segundo a teoria dos pensamentos de ordem superior, sugeriria consciência. Todavia poucas pessoas atribuem consciência ao seu portátil mas, por outro lado, intuitivamente atribuímos consciência fenoménica à maioria dos animais, sem que sintamos ser necessário que tenham uma percepção de ordem superior sobre os seus próprios estados mentais (Seager, 2004).

Um bom sumário das críticas às teorias de ordem superior pode ser encontrado em (Lycan, 2009).

6.2 Modelos baseados em teorias cognitivas

As teorias cognitivas tendem a associar a consciência a certas arquitecturas cognitivas ou específicos padrões de actividade no interior das estruturas. Searle (1992) entende que estudar os sistemas cognitivos é estudar a consciência, deixando patente a relação categórica entre os processos

[37] Cópia de um comando motor enviado aos músculos para produzir um movimento. A cópia não produz movimento, mas segue para outras regiões do cérebro informando-as da iminência do movimento.
[38] CERA (*Conscious and Emotional Reasoning Architecture*) CRANIUM (*Cognitive Robotics Architecture Neurologically Inspired*)

mentais como percepção, aprendizagem, inferência, tomada de decisões, resolução de problemas, emoções, etc., e a consciência. Das várias teorias cognitivas saliento:

6.2.1 Teoria do Espaço de Trabalho Global

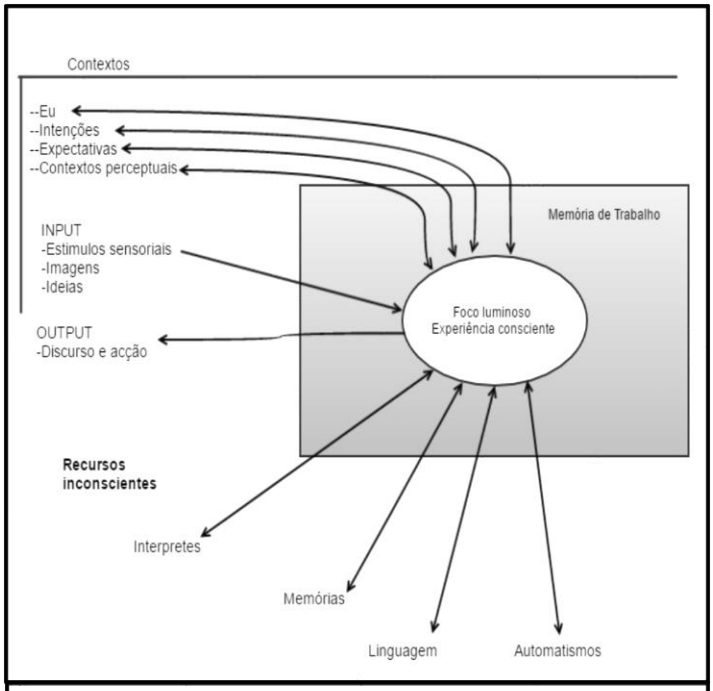

Figura 6-Esquema do Espaço de Trabalho Global. As setas dirigidas para dentro representam competição e as dirigidas para fora representam disseminação (adaptado de (Arrabales, 2011)).

A TETG, de **(**Baars, 1988), encara o cérebro humano como uma rede de processadores automáticos responsáveis pelas sensações, controle motor, linguagem, raciocínio, etc. A maior parte do processamento é inconsciente e ocorre em áreas específicas do cérebro, mas há um Espaço de Trabalho Global (ETG) distribuído, de capacidade limitada, cujos conteúdos podem ser difundidos aos diversos processadores especializados, para acesso e uso geral.

A consciência resultará da competição pelo ETG e é a informação difundida que é consciente.

Baars (1997) utiliza a metáfora de um teatro no qual a atenção é o foco de luz (fig. 6) que varre o cenário (a memória de trabalho). Neste os "actores" (conteúdos inconscientes) competem para aparecer sob o foco, e a selecção é feita nos bastidores[39] pelos processadores não conscientes, tendo em conta o contexto e os conjuntos de crenças (geralmente não conscientes) que determinam os pensamentos conscientes (a actuação na cena).

Para Baars a consciência é uma ferramenta para aceder aos conteúdos não conscientes da mente, e também o órgão de difusão do cérebro, que facilita o acesso a disseminação e a troca de informação, bem como coordenação global e controle. Alguma imagiologia parece validar partes importantes desta teoria (Baars et al., 2003).

Também (Dehaene et al., 1998) sugerem que a percepção consciente implica a activação da grande rede global, não sendo suficiente a actividade nas áreas de associação primária.

[39] O que sugere que, em grande medida, não temos acesso às razões pelas quais fazemos as coisas.

Outros, como Block (2007), entendem que a actividade recorrente local entre zonas baixas e altas do córtex sensorial pode bastar, mesmo na ausência de reportabilidade verbal e outros indicadores de consciência de acesso.

A TETG inspirou directamente o IDA (*Intelligent Distributed Agent*) (Baars & Franklin, 2007), um sistema multiagente da *USNavy*. Este e a sua mais recente versão, o *Learning* IDA, podem conversar em linguagem natural usando o correio electrónico, aceder a bases de dados, ajustar-se às políticas da marinha e comprovar os requisitos de cada trabalho, custo associado e preferências de cada marinheiro[40]. Para Franklin et al. (2012), o facto de os agentes se comunicarem entre si através de um ETG implica que o modelo é funcionalmente consciente

- Dehaene et al. (2003) desenvolveram um modelo de ETG para comparar padrões de actividade neuronal na execução de tarefas de rotina[41] e de tarefas conscientes. O modelo simula uma rede de regiões do córtex cerebral, cada uma das quais representa um processador especializado ou o ETG.

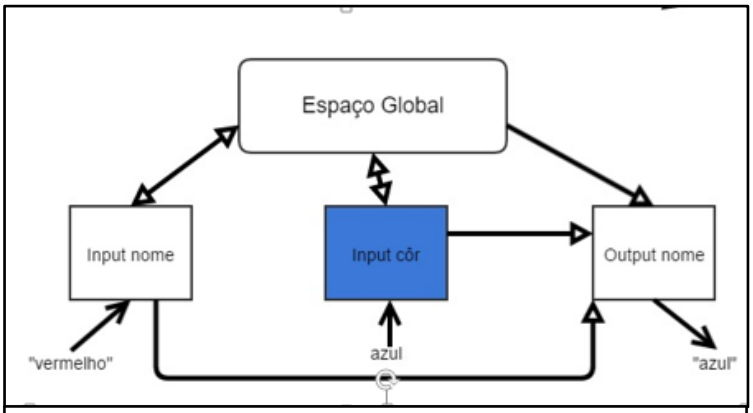

Figura 7- Esboço da arquitectura de um modelo ETG neuronal para a tarefa *stroop*. Cada rectângulo representa uma região cortical, constituída por um conjunto ou camada de neurónios excitatórios e inibitórios. Processos localizados ocorrem nas regiões de *input* e *ouput*. O processamento global é feito na camada do ETG. (adaptado de (Reggia, 2013)).

A versão inicial (fig. 7) foi concebida para simular o desempenho humano na tarefa *stroop*[42]. O modelo foi treinado em primeiro lugar para nomear as cores de entrada apresentadas isoladamente (tarefa fácil), depois para nomear palavras de entrada incongruentes com as cores (menos fácil) e, finalmente, para nomear cores de entrada com nomes incongruentes (tarefa difícil). Nas primeiras duas tarefas o ETG activou-se pouco. Já na tarefa difícil observou-se uma activação substancial e selectiva do ETG, à medida que os neurónios aprendiam a suprimir a actividade na região que manipula palavras de *input*, até que a tarefa se tornou rotineira. Os autores entendem que o processo corresponde ao que acontece com os seres humanos, nos quais há uma correlação entre a actividade generalizada do córtex cerebral e tarefas que exigem esforço consciente.

- Shanahan (2010) também entende que as acções inconscientes são automáticas, sem intervenção da atenção consciente, ao passo que as conscientes implicam reportabilidade introspectiva,

[40] Por exemplo, numa aplicação do IDA para escalar marinheiros para novas tarefas, através de um diálogo em linguagem natural, os processos reconhecem partes específicas do texto, categorizam-nas e contribuem com informações para o ETG.
[41] Que em seres humano não requerem esforço consciente.
[42] Tarefa de nomear palavras com cores congruentes e incongruentes.

flexibilidade em situações novas, capacidade mental para executar algoritmos e evocar memórias, etc. Para Shanahan há muitas coisas a acontecer simultaneamente no cérebro, e uma característica fundamental da condição consciente é a integração de tudo isso. O mecanismo da integração é o ETG, mas Shanahan entende que o *design* de Baars roça o *homunculus*[43], pelo que rejeita a metáfora do teatro e propõe que o ETG seja um interface que liga entre si as unidades especialistas (fig. 8). O conteúdo da consciência é a informação que flui pelo ETG.

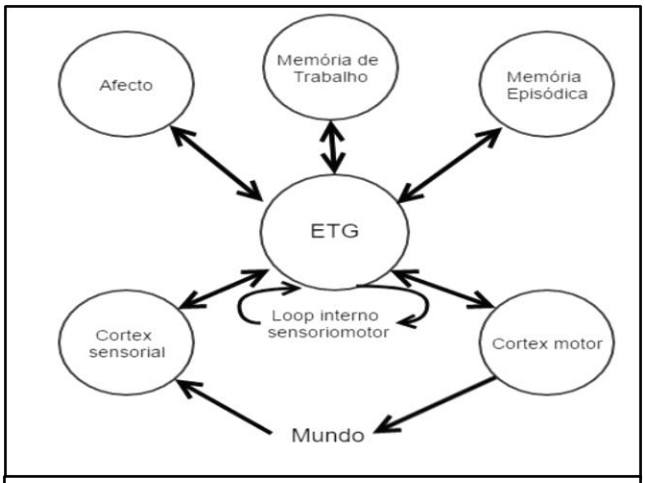

Figura 8- Modelo de Shanahan. O ETG liga os módulos que competem por ele. Um *loop* interno sensoriomotor permite desencadear ações a partir de estímulos mentais. (adaptado de Shanahan (2010)).

O modelo explica a integração e a capacidade de imaginar uma acção antes de executar *(loop)*, mas não explica como e porquê a suposta actividade neural consciente no ETG aparece internamente como *qualia*. Logo não vai ao encontro do problema duro.

Saliente-se que a maioria dos trabalhos que implementam a TETG visam expressamente a simulação da consciência funcional e não tanto a sua instanciação embora Raffone & Pantani (2010) sugiram essa possibilidade, até porque a ideia de que um ETG distribuído é um correlato neuronal da mente consciente, é consistente com as visões contemporâneas sobre a natureza distribuída da cognição numa rede estrutural e funcionalmente interconectada de regiões do córtex (Sporns, 2011).

Os modelos de ETG explicam parcialmente a unidade da consciência e corroboram a ideia de que a actividade cerebral consciente é globalmente distribuída, ao passo que não consciente é localizada. São também consistentes com a tese de que a limitada capacidade da memória de trabalho se deve à competição pela representação global. Podem até fazer previsões sobre anomalias expectáveis em máquinas conscientes (Wallace, 2006).

Quanto à consciência fenoménica, Raffone & Pantani (2010) sugerem que pode eventualmente emergir de interacções neuronais recorrentes, em *loops* cerebrais que ligam partes distantes mas conectadas do cérebro. Todavia a verdade é que estas teorias e estes modelos não explicam por que razão o processamento global de informação é um CNC, não dão detalhes sobre como funciona a maquinaria neuronal subjacente, nem explicam porquê e como a actividade neuronal aparece

[43] Conhecida falácia que consiste em imaginar um pequeno homem que, dentro do cérebro, coordena as acções do corpo. É uma falácia porque a questão se pode colocar recorrentemente para o próprio homúnculo, que terá também dentro dele um ainda mais pequeno, e assim sucessivamente, numa regressão infinita.

internamente como *qualia*, logo não abordam o problema duro e por isso não são suficientes para informar a instanciação da consciência na máquina ou seja, não é claro em que é que o ETG contribui especificamente para a consciência. Como diz Shanahan (2006), por agora a instanciação da consciência é ficção científica.

6.2.2 Implementações baseadas em mecanismos de atenção

Em cada momento um organismo presta atenção consciente apenas a uma pequena parte do fluxo de informação sensorial que lhe chega do mundo e do corpo. A evidência neurológica mostra que os mecanismos da atenção modulam a informação sensorial recebida, atenuando a actividade neuronal irrelevante e acentuando a actividade que representa os objectos tidos como importantes (Buschman & Miller, 2007). Sendo fenómenos diferentes, atenção e consciência aparecem normalmente ligadas e correlacionadas (Treisman, 2009), razão pela qual vários estudos de CA exploram a modelação de mecanismos de atenção.

Alguns investigadores consideram os mecanismos de atenção como sendo a base funcional da consciência e outros identificam um aspecto ou componente específico destes mecanismos como os responsáveis pela atenção consciente.

- Exemplificando a primeira abordagem, Tinsley (2008) descreve uma RNA cujo *output* é considerado a representação consciente de um estímulo. Trata-se de uma rede multicamada de regiões responsáveis pelas entradas sensoriais, integração, modulação, etc. As diferentes modalidades sensoriais são processadas em paralelo e convergem numa região selectora que produz o *output*. A modalidade sensorial e/ou a localização do estímulo podem ser usadas para seleccionar a parte do padrão de actividade de entrada que atinge em cada momento o *ouput* do modelo e entra na consciência. Embora a implementação seja bastante abstracta e limitada, ilustra bem como os mecanismos atencionais seleccionam o conteúdo das "representações conscientes" em redes estruturadas, embora não explique porquê, nem aborde a questão dos *qualia.*

- A Arquitectura Cognitiva de Haikonen (ACH), entre outras coisas enfatiza também os mecanismos de atenção. Haikonen (2007a), para quem a consciência é a presença de *qualia,* acredita que a CA poderá emergir em organismos autónomos que tenham uma arquitectura complexa, adequada e inspirada no funcionamento do cérebro. A ACH é um sistema neuronal sub-simbólico/simbólico, inspirado no cérebro e cognição humanos, embora não os procure modelar estritamente. No sistema a consciência é uma maneira de operar caracterizada por representações distribuídas de sinal, processos mistos de percepção[44], reporte intermodal e introspecção. É

[44] O modelo não representa internamente uma bola como um gráfico redondo, nem como um conjunto de números que indicam diâmetro, cor, etc. mas sim por um vector de sinais distribuídos. Se lhe mostram bolas de diferentes tamanhos, cores, etc., e a cada exposição o seu microfone capta um padrão de som (palavra "bola"), o modelo associa o padrão sonoro ao visual e é neste tipo de associações que se baseia o seu processo de percepção.

constituído por vários módulos intercomunicantes, semelhantes aos processadores não conscientes especializados da TETG, mas não existe um ETG separado. Um tópico é considerado consciente se o organismo como um todo lhe presta atenção, sendo esta caracterizada do seguinte modo: em cada momento, cada módulo especializado pode transmitir a todos os outros informações sobre o tópico que está a considerar. O tópico que ganha mais atenção colectiva entra na consciência e o organismo torna-se consciente (Haikonen P. , 2007a).

Segundo o autor, esta arquitectura cognitiva será capaz de percepcionar o seu próprio corpo e estado interno. Por exemplo, quando as mãos tocam no próprio corpo produzem-se dois conjuntos de sinais que permitem ao organismo deduzir que o que está a tocar é a sua própria "pele". De igual modo, num estado inicial do desenvolvimento cognitivo do organismo, quando as suas mãos se movem frente aos sensores visuais, descobrirá, tal comos os bebés, que são parte do seu próprio corpo.

Haikonen afirma que este modelo poderá ser capaz de gerar o fluxo de imagens mentais que caracteriza a experiência de seres conscientes. Por exemplo, quando surge o estímulo visual da bola amarela, emerge na máquina um padrão de sinais entre os quais há atributos activos como "redondez" e "amarelo". A activação deste padrão evoca outros padrões conhecidos, como o padrão de som da palavra "bola", o conhecimento da "bola rolando", a memória visual de uma "bola azul" que já tinha sido vista antes, etc. E todas estas representações podem evocar outras relacionadas: "rola mais depressa numa pendente", a imaginação da "bola rolando para cima", etc.

O modelo inclui também simulação de emoções, através de regras que constroem a capacidade de catalogar (mas não de sentir) o que é a dor, o prazer, bom e mau, agindo em conformidade[45], e é consistente com a evidência neurobiológica de que estados cerebrais conscientes estão associados à comunicação global entre as regiões do córtex cerebral (Massimini et al., 2005). Implementado parcialmente num *robot*[46] (Haikonen P. , 2012), a ACH sugere que o organismo só pode ser consciente de um tópico de cada vez e que pode estar autoconsciente se todos os seus módulos estão a trabalhar num tópico que envolve o próprio organismo.

-Quanto aos estudos que identificam um aspecto específico dos mecanismos de atenção, Kuipers (2008) propôs um modelo no qual a atenção selectiva será operacionalizada através de um apontador simbólico que, em cada momento, mantém correspondência entre um conceito de alto nível simbolicamente representado, e a sua representação de baixo nível no fluxo de dados em constante mudança. A parte do mecanismo de atenção que faz o "*symbol grounding*", ancorando representações simbólicas a específicos segmentos espaçotemporais do fluxo de dados sensoriais, é considerada

[45] Por exemplo: Medo: "mau"+ "dor" = fugir; Desejo: "bom" + "prazer" = aproximar.
[46] Não é totalmente implementável porque, segundo Haikonen, ainda não existe tecnologia capaz.

como responsável pela consciência. Embora nenhum sistema assim tenha sido ainda implementado, para Kuipers nada impede que as características essenciais da consciência possam, em princípio, ser instanciadas num *robot* integrado no seu ambiente, com suficiente poder computacional e um sistema sensoriomotor suficientemente rico. Todavia Chella & Gaglio (2012) argumentam que a redução de dimensionalidade implicada pelo uso de apontadores simbólicos é exatamente o oposto do que é necessário para instanciar consciência fenoménica.

- Nos últimos anos Taylor (2003a, 2007, 2012) aperfeiçoou o modelo CODAM[47], porque se entende que a *corollary discharge*[48], associada a mecanismos descendentes de controlo, é uma característica essencial da maquinaria neuronal subjacente à consciência (Cotterhill, 2003). Para Taylor (2007), é devido ao sinal de *corollary discharge* que o organismo tem a experiência consciente de se sentir responsável pelas acções que controlam a sua atenção. A ideia tem sido aplicada na criação de sistemas de controle neurocomputacionais para direccionar os movimentos de *robots* (Oh et al., 2012) e também na compreensão dos mecanismos cerebrais de controle motor.

No modelo CODAM (fig. 9) o sistema executivo (córtex pré-frontal dorso lateral) envia sinais de controle para um modelo inverso (córtex parietal) que modula selectivamente o *input* no córtex sensorial posterior, amplificando a informação que merece atenção e reduzindo a outra. Segundo Taylor o modelo inverso produz uma *corollary discharge* que informa um sistema de monitorização (giro cingulado e lobo parietal) das mudanças previstas no foco de atenção. Acredita-se que a *corollary discharge* redirecciona o foco de atenção e contribui para a correcção de erros. Em implementações na área da psicologia cognitiva alguns resultados têm sido qualitativamente semelhantes aos observados com sujeitos humanos (Taylor & Fragopanagos, 2007).

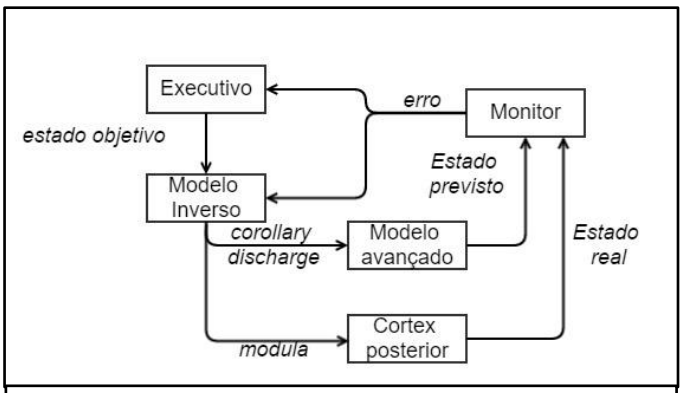

Figura 9- Modelo CODAM (simplificado) de controlo da atenção. (adaptado de (Taylor, 2007)).

Os modelos computacionais referidos, baseados em mecanismos de atenção, tratam, como vimos, do processamento da informação consciente. São plausíveis, dado o conhecido papel que mecanismos neurobiológicos da atenção desempenham na determinação dos conteúdos da consciência humana, e

[47] *Corollary Discharge of Attention Movement*
[48] Um estudo usando fMRI, MEG (Magnetoencefalografia) e EEG concluiu que há substanciais indícios da existência de um sinal de *corollary discharge* no cérebro humano (Taylor, 2012).

procuram relacionar a compreensão do controle da atenção com a natureza da consciência, mas tomemos por exemplo o organismo artificial de Haikonen.

Será consciente, no sentido que o próprio autor define como condição necessária isto é, experienciará *qualia*? É verdade que parece percepcionar o mundo e o seu corpo, e que o significado de cada percepto está alicerçado no mundo. É verdade que tem um fluxo de imagens no seu cérebro artificial e percepciona-o. Tem um mecanismo para simular as emoções e, uma vez que fala, pode supor-se com potencial para reportar o seu fluxo mental e a sua percepção dele. Tem comportamentos que parecem conscientes. Para Doan (2009b), se a consciência é uma questão de tudo ou nada, e o limiar não está ao nível de um adulto humano, o organismo de Haikonen é consciente. Se é um *continuum*, a aplicação da ConsScale (Arrabales, 2011) coloca-o a par de alguns animais a quem atribuímos consciência.

Todavia o modelo não parece ser capaz de deduzir que "eu" é ele mesmo, nem de saber que as percepções e emoções são os seus próprios *qualia*, como o próprio Haikonen (2012) reconhece.

Em suma, as teorias que equiparam a atenção à consciência não explicam de forma convincente por que razão a selecção de informação para ser representada como um padrão de actividade numa rede neuronal, a torna consciente, nem é claro um eventual papel causal na criação da experiência consciente.

6.3 Teoria da Integração da Informação

Tononi (2004, 2008) sugere que a consciência corresponde à capacidade de um sistema para integrar informação e, como vimos, a integração da informação pelo foco da atenção num tópico é também essencial na TETG (Baars, 1988) e no modelo de Shanahan (2010).

A TII baseia-se nos seguintes postulados (Tononi & Koch, 2015)

- A consciência existe em cada indivíduo e é independente de observadores externos.

- A consciência é estruturada e cada experiência é composta por muitas variantes fenoménicas[49].

- A consciência é específica e cada experiência é composta por um certo conjunto de específicas variantes fenoménicas, sendo necessariamente diferente de outras possíveis experiências que o organismo pode ter[50].

-A consciência é unificada e cada experiência é irredutível a quaisquer subconjuntos de outras coisas. Se, por exemplo, percebo um triângulo vermelho, a minha experiência não é constituída pela experiência de um triângulo sem cor mais a experiência de um vermelho sem forma.

-A consciência é definitiva no conteúdo e no grão espaçotemporal. Cada experiência tem o conjunto de variantes fenoménicas que tem, nem mais nem menos e flui à velocidade que flui.

[49] Na mesma experiência posso distinguir uma mesa, um caderno, uma caneta, um som, um cheiro,etc.
[50] Uma experiência de pura escuridão é o que é, porque, entre outras coisas, não contém objectos.

A TII faz previsões refutáveis e aplica uma medida matemática (φ) que mede a integração da informação nas partes de um sistema e na organização que integra essas partes, tentando assim quantificar a presença de consciência. Basicamente um qualquer sistema pode conter muitos módulos interactuantes e o que tiver um φ mais alto será mais consciente. Balduzzi & Tononi (2009) sugeriram que a quantidade de consciência é a quantidade de informação integrada e os *qualia* são especificados pelo conjunto de relações informacionais entre elementos da informação integrada. Tentam descrever geometricamente todo o conjunto de ligações da informação integrada, considerando um espaço de *qualia* com um eixo para cada possível estado do sistema. Deste modo cada *quale* pode, em princípio, ser geometricamente mapeado e encontrado um correspondente padrão de actividade neuronal. Hoel (2017) sugere que os agentes[51] emergem causalmente de certos estados macroscópicos de um sistema[52], pelo que consciência deverá ser estudada a partir desses estados macroscópicos que são, *grosso modo*, os que Tononi considera na TII.

Em termos metafísicos, para a TII a consciência é uma propriedade fundamental e intrínseca da realidade, que ocorre em diferentes graus, correspondentes ao valor de φ. Um mecanismo simples, que não esteja integrado num mais complexo, terá um certo grau de consciência, o que concorda com certos aspectos das teorias metafísicas pampsiquistas, como o próprio Tononi admite[53].

A teoria pretende dar conta, não só da quantidade mas também da qualidade de uma experiência individual[54] e contém um cálculo para avaliar se um determinado sistema físico é consciente e de quê. Rejeita o funcionalismo computacional e prediz que nem a mais complexa rede neuronal artificial, nem os mais sofisticados computadores digitais serão conscientes, mesmo que exibam um comportamento funcionalmente equivalente ao nosso. Tononi & Koch (2015) referem mesmo que as mais perfeitas simulações do cérebro humano, nada experienciarão. Respondem à experiência conceptual de Chalmers (2010), sugerindo que uma simulação digital, neurónio a neurónio, sinapse a sinapse, de um cérebro humano, não experienciará nada, mesmo que o seu comportamento seja indistinguível do original. Será um *zombie* e nenhum sofisticado TT servirá para demonstrar a presença de consciência[55].

A TII tem pouco mais de 10 anos e é bastante recente o desenvolvimento de modelos neuronais de consciência nela inspirados. Todavia a medição de φ em grandes redes é um problema de enorme complexidade temporal, o que coloca barreiras práticas à implementação. Na verdade só tem sido possível usá-lo em RNA muito elementares.

[51] Entidades com comportamento orientado por intenções e objectivos
[52] E não das interacções clássicas dos componentes micro (físicas, químicas e celulares) o que coloca em causa a estrita ideia reducionista.
[53] Embora não subscreva a ideia de que tudo é consciente.
[54] A qualidade da consciência em diferentes partes do sistema é determinada pelas relações informacionais (Tononi, 2004).
[55] Shanahan (2015) responde a esta experiência conceptual com uma outra em que o cérebro substituído é o do próprio Tononi, concluindo que é precipitado e prematuro "decidir" o fim do funcionalismo.

Um estudo de Aleksander & Gamez (2009) usou redes neuronais dinâmicas[56] para memorizar imagens, examinando como diferentes padrões de conexão e distintas arquitecturas influenciam a integração de informação numa rede e verificaram que é maximizada pela forte conectividade distribuída na rede.

Num estudo subsequente, φ foi sistematicamente calculado para redes simples de quatro neurónios e comparado à "*liveliness*"[57] da rede (Aleksander & Gamez, 2011). O trabalho mostrou que a *liveliness* da rede está fortemente correlacionada com φ, em redes pequenas.

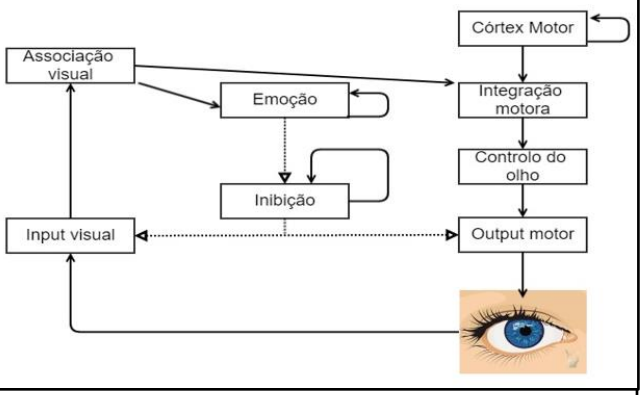

Figura 10- Esquema da arquitectura do neurocontrolador. As caixas representam camadas/regiões de neurónios simulados e as setas representam as conexões. As conexões pontilhadas são inibitórias. (adaptada de (Gamez, 2010)).

Tononi (2008) alega que as partes de uma rede neuronal associadas a elevados valores φ podem ter consciência fenoménica. A partir dessa hipótese Gamez (2010) desenvolveu um neurocontrolador (fig. 10) para um sistema robótico de visão. O objectivo era aprender a olhar preferencialmente para blocos vermelhos e "fugir" de blocos azuis. A direcção do olhar robótico é determinada por um padrão de actividade na região do córtex motor que actua através de um sistema multicamadas de controlo motor. O sistema aprende, de forma competitiva *hebbiana*[58], a associar padrões de actividade na região de integração motora com objectos no ambiente, alterando as ligações entre as regiões de integração motora e associação visual. Uma vez treinados, os mecanismos incluídos (não mostrados na figura) levam o sistema a olhar na direcção dos blocos vermelhos. As camadas de emoção e inibição filtram a actividade de outras regiões. A activação da região de inibição fecha o *input* visual e o *output* motor, isolando o controlador do *robot* físico. Quando o controlador está *offline* os padrões aleatórios de actividade que aparecem no córtex motor geram miradas do olho. Quando este incide num bloco vermelho a emoção é ativada e desligada a inibição, reconectando o controlador com o ambiente. Após o treino foi examinada a integração da informação nas regiões da rede para avaliar se as partes individuais do sistema poderiam ser consideradas conscientes. Como a aplicação directa do algoritmo de Tononi & Sporns (2003) é computacionalmente intratável para uma rede desta dimensão[59], Gamez aplicou um algoritmo aproximado e validado. A análise atribuiu o máximo φ a uma sub-rede com 91 neurónios que inclui toda a região de inibição, a maior parte da

[56] Que tendem a evoluir para um padrão estável (atractoras).
[57] Medida da influência causal entre 2 neurónios para um estado particular da rede. Se o neurónio 1 se liga com o neurónio 2 com uma conexão "viva", há uma maior probabilidade de que uma mudança no estado de 1 leve a mudança no estado de 2, independentemente de outras conexões.
[58] Aprendizagem associativa na qual a activação simultânea de neurónios faz aumentar o peso das ligações.
[59] Exigiria mais de 100 000 anos, com os actuais recursos computacionais (Reggia, 2013)

região de emoção e alguns neurónios de outras regiões. Análises posteriores calcularam a "consciência prevista por neurónio" e sugerem que, nesta rede específica, apenas as regiões de emoção e inibição estariam significactivamente correlacionadas com a consciência.

Este resultado é surpreendente porque os componentes com elevado φ são essencialmente os circuitos "*gating*" (inibitórios). Em arquitecturas neuronais o *gating* tem sido crescentemente reconhecido como muito importante para os mecanismos de controlo cognitivo da memória de trabalho, algo que alguns consideram ter uma estreita associação com a mente consciente, e que tem sido o foco de estudos computacionais recentes (Sylvester et al., 2013).

Apesar da sua crescente popularidade, a TII enfrenta importantes barreiras filosóficas e práticas. Do ponto de vista filosófico está muito longe de ser consensual a ideia de que a experiência subjectiva seja equivalente à capacidade de integrar informação, e por isso não é claro que φ tenha algo a ver com a experiência subjectiva que associamos à mente consciente (Manzotti, 2012). Foi, por exemplo, demonstrado que escolhendo adequadamente os pesos sinápticos se podem construir redes neuronais simples, totalmente ligadas, com valores φ arbitrariamente altos (Seth et al., 2006). Ora, de acordo com a TII, a partir de um certo φ (não especificado) tais redes teriam consciência fenoménica, conclusão que intuitivamente se recusa e que, de qualquer forma, não clarifica a natureza da consciência (Seth, 2009).

Quanto às barreiras práticas, não é temporalmente possível computar directamente o φ em redes neuronais de grandes dimensões e, mesmo que fosse, o valor φ, por si só, não fornece uma indicação significativa da presença de consciência, por falta de um valor de referência. A referência adequada seria obviamente o φ de um cérebro humano, medida fora do nosso alcance, não só pela dimensão das redes envolvidas, mas também porque não conhecemos assim tão bem a estrutura funcional do cérebro humano. Está a ser feito um esforço para desenvolver algoritmos computacionalmente eficientes que possam ser aplicados a modelos cerebrais de maior escala actualmente em desenvolvimento[60] (Cattel & Parker, 2012) os quais, apesar de lidarem já com centenas de milhares de neurónios e ligações, estão longe de exibirem *performances* comparáveis sequer aos mais simples cérebros biológicos.

Recentemente Tegmark (2014), numa aproximação que tenta generalizar a TII a sistemas quânticos arbitrários, sugeriu a hipótese de que a consciência possa ser um estado da matéria, equiparável aos tradicionais e conhecidos estados físicos.

[60] *SpiNNaker* (Universidade de Manchester), *Blue Brain* (Escola Politécnica Federal de Lausana), *C2S2 SyNAPSE* (IBM), *FACETS* (Universidade de Heidelberg) *Neurogrid* (Stanford), *IFAT e o NeuroDyn* (Universidade da Califórnia).

6.4 Teorias neuronais e Correlatos Neuronais da Consciência

Segundo o enfoque estritamente fisicalista "o nosso sentido de identidade pessoal e de livre arbítrio, não são, na verdade, senão a conduta de vastos conjuntos de neurónios e moléculas associadas" (Crick, 1994). Embora pouca gente concorde hoje com esta visão tão redutora, prevalece a ideia de que a actividade das redes neuronais do cérebro gera os processos mentais, e que certos estados funcionais levam à consciência como uma propriedade emergente de massivas e paralelas computações. A pesquisa na CA insere-se também numa perspectiva funcionalista, procurando emular artificialmente os processos neuronais correlacionados com os processos mentais conscientes.

A maioria dos estudos neurobiológicos distribui-se pelos níveis neuronal[61] e de redes ou grupos de neurónios[62], e visa captar a diferença entre os processos mentais que se correlacionam com a consciência e os outros, bem como especificidades dos neurónios implicados nos processos conscientes, e na forma como se ligam e activam (Crick & Koch, 2003)[63].

A hipótese funcionalista dos CNC assenta no facto de haver nos seres humanos muitas redes neuronais com funções específicas, pelo que é tentador especular que existem também determinadas estruturas correlacionadas com a consciência.

Segundo Koch C. et al (2016), estar consciente significa experienciar e só não se está consciente quando se está num sono sem sonhos ou sob anestesia geral. Estar consciente parece estar correlacionado com algumas funções executadas pelo cérebro, quer seja pela sua organização funcional, quer seja pelas peculiaridades do material de que é feito, e testar isto implica fazer experiências que alterem o material sem alterar as funções e vice-versa. Há já quem esteja a trabalhar em próteses de silicone para várias partes do cérebro[64], pelo que é de esperar que, mais tarde ou mais cedo, possam ser extraídas conclusões sobre a eventual influência do material de que é feito o cérebro, sobre os fenómenos mentais, como a memória, consciência, etc. De qualquer maneira, considerando que não é incomum que pessoas com lesões cerebrais confabulem e se enganem quanto às suas experiências, como acontece, por exemplo, na anosognosia visual[65], talvez nem essas conclusões sejam definitivas. Estudos sobre *blindsight* mostram que certas lesões no cérebro podem cancelar a experiência consciente da visão sem abolir a visão em si mesma (Lindsay et al., 2015). Da mesma maneira, aparentemente a consciência pode ser mantida na ausência de partes importantes do cérebro, como o comprova o caso da mulher que nasceu sem cerebelo (Thomson, 2014) mas estava completamente consciente, tinha experiências subjectivas e estava apta a descrevê-las. Por outro lado,

[61] Neste nível estuda-se o funcionamento concreto de cada célula cerebral e dos processos nas sinapses.
[62] Neste nível o estudo centra-se tanto nos pequenos grupos funcionais de neurónios, como nos grandes conjuntos como, por exemplo, a rede de percepção visual (Crick, 1994).
[63] Há quem considere (O'Reagan, 2007) que a procura dos CNC é um erro conceptual, sugerindo que a consciência P é um fenómeno que se manifesta durante a interacção do indivíduo com o seu ambiente, não sendo possível defini-lo como uma propriedade estrutural ou funcional do cérebro.
[64] Por exemplo, de um hipocampo artificial (Hampson et al.,2013)
[65] Os pacientes estão corticalmente cegos, mas afirmam peremptoriamente ser capazes de ver.

na doença de Alzheimer pode-se perder o sentido do passado e da identidade, a noção de existir e a própria representação do organismo (Damásio A. , 2003).

Sabe-se hoje que as bases anatómicas dos CNC estão principalmente localizadas na região temporal-parietal-ocipital, com contribuições de algumas regiões anteriores[66](Koch et al., 2016). A formação reticular do tronco cerebral, tálamo e partes do córtex postero-medial possibilitam interacções entre áreas corticais que contribuem directamente para o conteúdo da consciência.

Um conjunto de ilustres personalidades na área do estudo da mente e da consciência, que inclui Norman Cook, António Damásio, Gil Carvalho, Harry Hunt e Oliver Sacks, em carta aberta a Cristof Koch (Cook et al., 2015) propuseram a tese de que o problema duro da consciência pode ter a ver com a resposta de células excitáveis aos estímulos externos que afectam a sua homeostase, sendo que a consciência de animais mais complexos poderia ser, em síntese, uma consequência da irritabilidade[67] coordenada e sincronizada de biliões de células excitáveis.

Na resposta, Koch referiu que embora o afluxo súbito de catiões seja necessário, não é suficiente, uma vez que, por exemplo, durante um ataque epiléptico os neurónios do córtex disparam de forma hiper-sincronizada, há um massivo afluxo de catiões para dentro das células e apesar disso o paciente perde rapidamente a consciência. Além disso a perda de grandes regiões dos gânglios basais e do cerebelo, que contêm 4/5 dos neurónios do cérebro humano, não tem qualquer impacto mensurável no conteúdo da consciência, como seria de esperar se a ideia de Cook et al. (2015) estivesse certa.

Como vimos, uma das propriedades fundamentais da consciência é a integração de múltiplas fontes sensoriais ou internas num objecto fenoménico e saber como isso acontece é conhecido pelo *binding problem*. Crick & Koch (2005) sugerem que a resposta pode passar pelo *claustrum*[68], que tem conexões de e para quase todas as áreas do córtex, funcionando como maestro da uma orquestra, coordenando as actividades de todos os componentes corticais. O trabalho de Koubeissi et al. (2014)[69] dá alguma sustentabilidade a esta tese.

Outro fenómeno apontado como solução para o *binding problem* é o padrão eléctrico conhecido por sincronia *gamma* (40 – 100 Hz)[70], que pode ocorrer localmente num indivíduo inconsciente, mas é mais espalhada no córtex cerebral em vigília. A hipótese de que a sincronia *gamma* possa ser um correlato da consciência e da experiência subjectiva (Ward, 2011) advém do facto de ser mais forte quando estamos conscientemente empenhados numa tarefa, e desaparecer quando estamos num sono

[66] Por exemplo em humanos, a activação das regiões ligadas ao reconhecimento facial está, forte e sistematicamente, correlacionada com ver faces, a sua estimulação pode induzir ou alterar a percepção das faces (Rangarajan et al., 2014), e lesões nessas áreas impossibilitam reconhecer faces familiares.
[67] Segundo os autores esta irritabilidade é desencadeada por mecanismos comuns a protozoários e a células de receptores sensoriais, neurónios e células musculares de organismos complexos. Estes mecanismos consistem, basicamente, na abertura e fecho de canais na membrana, propiciando a súbita passagem de iões.
[68] Estruturas finas e laminadas de neurónios, anichadas sob o córtex cerebral. Uma em cada hemisfério.
[69] Implantaram um eléctrodo perto do *claustrum* de uma paciente com epilepsia e, quando ligado, a paciente ficou inconsciente. Assim que foi desligado a consciência voltou e o processo repetiu-se, sempre com os mesmos resultados.
[70] Sincronia é a activação instantânea de muitos neurónios relacionados entre si. Crick & Koch (1990) sugeriram que a activação sincronizada a 40 Hz de conjuntos de neurónios era a base física da consciência mas acabaram por se retratar (Crick & Koch, 2003)

sem sonhos. Todavia Koch, et al. (2016) são da opinião que tanto a actividade *gamma*, como a sincronia são ilusórias assinaturas da consciência.

Em vários estudos verifica-se que todo o cérebro está interconectado e que a informação flui complexa e rapidamente entre regiões, mas é praticamente consensual entre os investigadores que o o tálamo é uma espécie de centro de comunicações entre as várias áreas do córtex, havendo uma forte correlação entre o complexo talamocortical e a presença de consciência (Ward, 2011).

Há também sugestões de que a rede entre o tronco cerebral, a ínsula esquerda ventral anterior e o córtex cingulado anterior pregenual, tem um papel importante na manutenção da consciência humana (Fischer, et al., 2016).

A imagiologia mostra também que a actividade metabólica em certas regiões corticais (córtex pré-frontal e parietal) baixa de forma significativa em coma e anestesia geral (Baars et al.,2003) e uma actividade metabólica mais globalmente distribuída e maior comunicação entre várias regiões corticais estão associadas com a aprendizagem consciente de novas tarefas (Baars, 2002).

Um outro possível correlato é a junção temporal parietal, já que activar esta área com TMS produz efeitos intrigantes como experiências fora do corpo e outras (Van Lommel et al., 2001).

Há todavia ainda muita incerteza acerca dos critérios a usar para determinar se uma determinada zona de cérebro participa na manifestação da consciência. Monti, et al. (2010) demonstraram que mesmo pacientes que se acreditava estarem em estado vegetativo, estavam na realidade em estado de mínima consciência ou *locked in* e Lamme (2006) e Block (2007) sugerem que a actividade local recorrente entre áreas altas e baixas do córtex sensorial pode ser condição suficiente para a consciência fenoménica, mesmo que não haja reportabilidade verbal ou indicadores da presença de consciência de acesso. Para Tononi (2012), o grande número de interacções causais no cérebro em conjunto com a natureza fugaz de muitas experiências, desafia os mais sofisticadas aproximações experimentais aos CNC.

Há muitas teorias neuronais[71] debruçando-se sobre diferentes processos cerebrais mas nenhuma é consensual e totalmente explicativa.

6.5 Teorias Quânticas

Na explicação clássica do mundo, a realidade física existe independentemente do observador e é regida por leis exactas e deterministas. Os nossos corpos e cérebros são parte desse mundo físico e regem-se pelas mesmas leis, mesmo que sintamos ter livre arbítrio para alterar o nosso

[71] Que vão desde modelos que enfatizam os campos globais integrados (Kinsbourne, 1988), *binding* através de oscilações síncronas (Singer, 1999), (Crick e Koch, 1990), padrões talamicamente modulados de activação cortical (Llinas, 2001), *loops* corticais reentrantes (Edelman, 1989), processos interpretativos baseados no hemisfério dominante (Gazzaniga, 1988), processos emocionais homeostáticos, somas sensoriais baseados no *nexus* frontal-límbico (Damasio,1999), etc.

comportamento. Todavia o fenómeno da consciência não se deixa facilmente capturar na descrição que a física clássica faz da realidade.

Será necessária uma descrição quântica? Esta é o mais básico nível de descrição que nos é possível no estado actual da Física, e há teorias que situam nele uma possível abordagem da consciência fenoménica. Uma delas é a *Orch-OR* (*Orchestrated Objective Reduction*) (Hameroff & Penrose, 2014) que ambiciona tratar directamente o problema duro. Para os autores, existe um ingrediente não algorítmico essencial nos processos conscientes, e a experiência consciente pode estar intrinsecamente ligada à estrutura fina da geometria espaçotemporal do universo (o que a coloca no âmbito da metafísica pampsiquista), sendo instanciada nos organismos por uma forma especial de colapso da função de onda quântica[72] nos microtúbulos do citoesqueleto de células cerebrais, que acontece sem a intervenção de um observador[73]. Segundo a teoria as tubulinas (proteínas que formam os microtúbulos) estão associadas a eventos quânticos internos e interagem com outras tubulinas. A superposição coerente macroscópica dos estados das tubulinas emparelhadas quânticamente tem lugar ao largo de áreas cerebrais de tamanho significativo, proporcionando a unidade global requerida pela consciência. Pizzi et al. (2010) verificaram de facto a existência de ressonância electromagnética nos microtúbulos, com uma frequência de 1510 MHz, sugerindo que com adequada instrumentação se podia provar a existência de frequências da ordem dos GHz e maiores. Hameroff et al. (2014) afirmam que os microtúbulos têm ressonância quânticas em frequências fractais[74] coerentes, num contínuo que vai de KHz a THz.

A teoria tem sido muito escrutinada mas sem relevantes aplicações práticas dada a dificuldade dos ambientes informáticos convencionais em lidarem com a mecânica quântica e particularmente com uma interpretação menos ortodoxa, como é o caso da *Orch-OR*.

6.6 Teoria do holofluxo

Esta teoria pampsiquista (Joye, 2016) junta os conceitos de cérebro holográfico[75] de Pribram (2013) e de *Implicate Order*, de Bohm (1980, 1990). Para este, o universo, que designa como o "Todo", consiste numa "*explicate order*", que é a realidade espaçotemporal que percebemos e

[72] A função de onda (ψ) é a solução da equação de Schrödinger e descreve o mais completamente possível o estado quântico de um sistema de uma ou mais partículas. Contém todas as informações sobre o sistema considerado isolado mas não é uma onda no espaço físico. Quando se dá o seu colapso (redução do vector estado) apenas um estado se "materializa", isto na estrita interpretação de Copenhaga.

[73] Segundo os autores a emergência da coerência quântica nos microtúbulos corresponde ao processamento pré-consciente (até 500 milissegundos), até que a diferença massa-energia entre os diferentes estados das tubulinas atinge um certo limite. Cada um dos estados superpostos tem as suas próprias geometrias espaçotemporais. Quando o grau de diferença de massa-energia coerente leva a uma separação suficiente da geometria espaçotemporal, o sistema tem de decidir entre reduzir-se ou colapsar num único estado. Produz-se assim uma superposição temporal de geometrias ligeiramente diferentes até se dar uma abrupta e clássica redução quântica.

[74] O padrão de informação repete-se nas diferentes escalas espaçotemporais.

[75] Descreve o cérebro como uma rede holográfica de armazenamento, em que cada parte contém toda a informação armazenada.

captamos e numa "*implicate order*" [76], não local, não espacial e não temporal, dobrada em dimensões adicionais, previstas pela teoria das cordas. Segundo Bohm a consciência não existe na estrutura do espaço-tempo, mas sim na "*implicate order*" e manifesta-se como uma energia de padrão holográfico que flui, em dimensões subquânticas, da *implicate order* para a *explicate order*. Segundo Bohm, o Todo está constantemente a dobrar e desdobrar entre as duas ordens.

Trata-se de uma teoria bastante inortodoxa mas coerente e bem sustentada, que dá boas respostas a fenómenos como, por exemplo a não – localidade[77] e as alucinações e experiências incomuns relatadas, entre outros, por Sacks (2012) em "*Hallucinations*".

Pylkkänen (2007) sugere que, segundo a interpretação de Bohm, cada partícula material contém informação de todo o universo, como acontece com os hologramas[78]. Bohm argumenta que a consciência e a matéria não estão absolutamente separadas, como sugere o dualismo cartesiano, mas formam um *continuum*, o Todo (Bohm & Hiley, 1993).

Segundo Joye (2016), o fluxo da consciência pode ser visto como um tipo de energia de plasma, veiculado em diferentes tecidos biológicos, cujas dimensões permitem frequências na ordem dos GHz, nomeadamente os microtúbulos.

6.7 Aplicações baseadas na teoria das emoções, sentimentos e consciência e na existência de um automodelo.

Visto que a experiência humana consciente envolve emoções, serão elas necessárias para a consciência?

As emoções são essencialmente factores motivacionais que levam a agir desta ou daquela maneira. Um computador digital não precisa de motivação, faz aquilo para que está programado, mas um organismo cognitivo, natural ou artificial, que se mova num mundo complexo tem vantagens em possuir um mecanismo que, face ao contexto, adapte o seu comportamento ou até a sua arquitectura as seus fins. Picard (1997) sublinhou a importância do desenvolvimento de sistema artificiais capazes de reconhecer, interpretar, processar e simular os sentimentos humanos (*affective computing)*, referindo que sem emoções as máquinas podem não ser capazes de comportamento criativo e inteligente. Alertou também para a possibilidade de, com emoção a mais, poderem eliminar o seu criador, mas Minsky (2006) defendeu que as emoções não são essencialmente diferentes do processo que designamos por "pensar".

[76] Bohm entende que a teoria quântica opera apenas num mundo cujo grão de tempo e espaço é superior aos limites de Plank (1,616 x 10-^{33} cm e 5.39 × 10^{-44} s), e que é abaixo desses limites, nos quais as leis da física falham, que se encontra a *implicate order*, esta sim, com verdeiro significado ontológico.
[77] A ideia de que a consciência nem sempre coincide com o funcionamento do cérebro e pode haver experiências fora do corpo.
[78] Os quais, como se sabe, têm a característica de cada uma das suas partes possuir a informação do todo

Algumas teorias enfatizam a existência de um modelo interno do próprio organismo e ligam a consciência aos sentimentos e emoções (Ciompi, 2003), na medida em que consideram que o estado emocional influencia fortemente a percepção e o comportamento.

Para Damásio (2011) a consciência é um particular estado mental no qual há conhecimento da existência própria (*self*) e do ambiente exterior. Um sentimento primordial e fundamental na construção do *self*, é o de que o próprio corpo existe e está presente. O *self* existe apenas na mente (não é o corpo nem existe fora dele) e emerge do processo de se observar a si mesmo, em várias fases, cada uma delas correlacionada com específicas partes do cérebro: Na primeira fase o proto *self* emerge da parte do cérebro que representa o organismo e consiste num conjunto de imagens que descrevem estados estáveis do corpo e geram sentimentos primordiais; Na segunda o cérebro cria um modelo de segunda ordem, o *core-self*, que representa o corpo, o mundo exterior (tal como o percebem os sentidos) e as interacções entre eles. Em cada instante o cérebro tem portanto duas perspectivas simultâneas: uma desde o exterior (o cérebro vê o corpo que tem um mão que agarra um chávena), e outra do modelo do corpo (que diz ao cérebro que os seus olhos veem a chávena). No cérebro estas duas vistas coincidem, pelo que assume que esse corpo é o seu corpo (sou eu que estou a agarrar esta chávena). O organismo torna-se consciente de si mesmo e alcança a *core consciousness*; Na terceira fase emerge um *self* autobiográfico resultante da interacção de memórias ou futuros antecipados com o proto *self*, dando origem a vários pulsos de *core self.*

Damásio entende que as imagens são autorreferências das disposições (memórias), que os sentimentos são autorreferências das emoções, que o *self* é uma autorreferência da mente e que a consciência é uma autorreferência da não consciência. Em suma, vivemos num *loop* de autorreferência e um organismo que tem emoções, história, é autoconsciente e recorda o passado, é consciente de si mesmo como entidade com uma história (Bosse et al., 2008).

Segundo Barron & Klein (2016) os insectos, embora com estruturas diferentes das dos mamíferos, têm também essa capacidade de criar um automodelo, condição que consideram suficiente para a experiência subjectiva, embora tanto a premissa como a conclusão sejam disputadas (Key et al., 2016). Recentemente demonstrou-se que até pacientes aparentemente não conscientes podem simular internamente a realização de diversas actividades (Monti, et al., 2010). De acordo com esta perspectiva qualquer organismo com sensores internos, estrutura apropriada e adequado *software* poderá gerar um modelo do corpo (Doan, 2009b), *ergo* uma *core consciousness.*

Várias implementações incorporam explícita ou implicitamente um automodelo. Metzinger (2000a) sugere que a mente inclui um automodelo resultante de uma representação interna das propriedades espaciais do corpo e argumenta que a experiência subjectiva e consciente de um "eu" surge porque a imagem corporal está sempre lá devido à entrada proprioceptiva, somática e sensorial

que o cérebro recebe constantemente, e devido à incapacidade cognitiva de reconhecer que a imagem corporal é uma representação virtual do corpo físico. Em suma a automodelação do corpo como um organismo que causa o comportamento, levaria à experiência subjectiva.

Nos últimos anos tem-se avançado a hipótese de a consciência fenoménica poder emergir dos mecanismos de raciocínio introspectivos sobre a percepção (Mc Dermott, 2007). Tem sido também sugerido que algumas propriedades essenciais de um conceito de si mesmo, como a existência, a continuidade temporal, etc., podem ser a base para a automodelação em organismos inteligentes, tenham ou não corpo (Samsonovich & Ascoli, 2005). Por conseguinte a introdução de um conceito de si mesmo durante a aprendizagem seria crucial na criação da "consciência computacional" em arquitecturas cognitivas adequadas (Samsonovich & Dejong, 2005).

Um tipo específico de modelo interno que se acredita poder capturar várias características da consciência é a arquitectura da máquina virtual (Sloman & Chrisley, 2003)[79]. Esta abordagem vê a mente humana consciente como um *software* executado no cérebro (*hardware)*. A automodelação ocorre quando o *software* desenvolve conceitos para categorizar os seus próprios estados tal como são detectados por monitores internos. Neste contexto, os *qualia* são um efeito colateral do facto de a máquina ter um componente que lhe permite examinar o seu próprio funcionamento e representações internas. Reconheça-se contudo que esta caracterização dos *qualia* é controversa e não lida directamente com o problema duro, tal como o reconhecemos.

Quanto a máquinas que mostrem capacidades emocionais ou que sejam capazes de as simular, há numerosos projectos e aplicações[80], que vão de sistemas que promovem gentileza e gratidão, a reconhecimento de *stress,* passando por avaliação em tempo real de comportamentos, ideação suicida, etc.

No contexto da robótica têm também sido feitos vários estudos que envolvem automodelação e emoções, dos quais particularizo:

- O *robot* CRONOS, projecto explicitamente focado na CA, inclui um automodelo que interage com um modelo interno do mundo processando informações sensoriais e movendo-se como o *robot* físico, tendo como referência o modelo do mundo. Argumenta-se que este processo mimetiza qualitativamente os conteúdos cognitivos da consciência humana, pelo que, para Holland (2007), pode ter consciência fenoménica. A intenção do projecto era, no final, examinar todos os sistemas, procurar sinais de consciência e descrever a sua fenomenologia, mas foi encerrado em 2007, sem nada de significativo se apurar relativamente à consciência fenoménica. Gamez, um dos maiores

[79] Uma máquina virtual é uma simulação de uma máquina, implementada em *software*, que corre numa máquina física diferente (*hardware*).
[80] *Projects in Affective Computing*, http://affect.media.mit.edu/projects.php.

dinamizadores, parece ter evoluído para algum cepticismo quanto ao poder da ciência nesta matéria, como mostra no seu último livro (Gamez D. , 2007).

- Ainda no campo da robótica, o trabalho de Takeno (2008, 2013) estuda *robots* "conscientes" controlados por redes neuronais recorrentes com aprendizagem supervisionada. O próximo estado previsto e o estado actual do mundo são comparados a cada momento e, quando coincidem, estipula-se que o *robot* instanciou a consciência já que, para Takeno, o que prova a consciência é a consistência entre a cognição e o comportamento. Estes *robots* mostraram capacidade de autorreconhecimento em espelho.

- Para estudar a utilidade do automodelo numa perspectiva evolucionista Bongard et al. (2006) utilizaram um *robot* físico com quatro pernas multiarticuladas, cada uma delas com sensores de toque e posição. Inicialmente o *robot* não tem informações sobre a dimensão das pernas ou sobre como elas estão ligadas ao corpo, e também não as sabe coordenar. Por tentativa e erro cria uma imagem do seu corpo e dos seus movimentos. Passa por repetidos ciclos de hipóteses e experimentação para construir a sua imagem corporal (fig. 11), começando com uma acção aleatória.

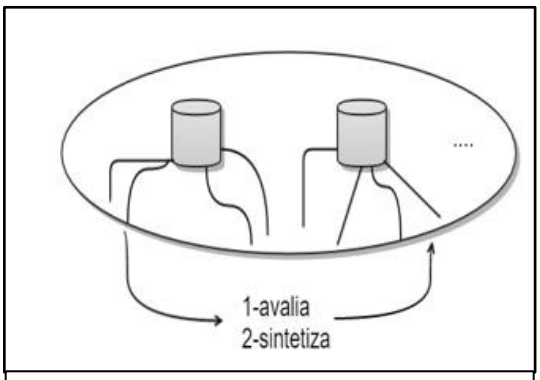

Figura 11- O agente mantém várias auto-imagens, repetidamente refinadas, por avaliação e correcção. (adaptado de (Bongard et al., 2006))

O *robot* "imagina" os resultados sensoriais esperados com diferentes acções, e escolhe depois a acção tida como adequada. Executa-a, recebe *feedback* sensorial e, comparando os resultados previstos com os reais, constrói um novo conjunto de automodelos candidatos. Após vários ciclos é avaliada a capacidade efetiva de movimento, gerando uma sequência de locomoção que é executada.

O automodelo permite que o organismo crie expectativas sobre o futuro, capacidade que alguns consideram estar associada à consciência (Ascoli, 2005). Os períodos em que, estando *offline*, o *robot* "imagina" as consequências das acções no seu automodelo, foram até interpretados como "sonhos robóticos" (Adami, 2006). Holland et al. (2007) sugerem que este tipo de automodelação é um passo importante para alcançar a consciência.

Quanto à utilidade parece confirmar-se que um organismo com automodelo tem vantagens (Hart & Scassellati, 2011), por se ter demonstrado que permite ao dispositivo robótico restaurar a funcionalidade após danos repentinos e inesperados. Esta hipótese é consistente com a tese de Dawkins (1976), segundo a qual a evolução da capacidade de simulação do cérebro pode ter culminado na consciência quando a simulação do mundo se tornou suficientemente complexa para incorporar um automodelo.

Até à data os automodelos desenvolvidos são bastante simples e lidam com baixos níveis de processamento sensorial e motor que não capturam a riqueza do metaconhecimento humano sobre outros aspectos do "eu", pelo que é expectável a extensão da automodelação a processos cognitivos mais complexos (Goertzel, 2011).

6.8 Projectos potencialmente relevantes

Embora não tendo explicitamente como objectivo a pesquisa em CA, há várias outras investigações e projectos que, numa perspectiva funcionalista, importa manter no radar. Trata-se, em alguns casos, de projectos complexos e dispendiosos, nos quais se aplicam as mais avançadas capacidades tecnológicas e humanas, e pode muito bem acontecer que, para além dos seus objectivos explícitos, surjam "efeitos colaterais" relevantes para o apuramento das possibilidades de instanciar a consciência em organismos artificiais. A história do conhecimento está repleta de avanços inesperados surgidos no âmbito de projectos que não os tinham como meta, como a *internet*, a penicilina, a radiação, a fissão nuclear, o raio X, etc.

6.8.1 Google Brain

O ***Google Brain*** tem interesse pela dimensão e complexidade das redes neuronais envolvidas. Foi lançado pela *Google* em 2011, tendo em vista desenvolver um sistema de *deep learning* em grande escala (Dean, et al., 2012), para treinar uma rede com biliões de parâmetros usando milhares de unidades de processamento. O sistema treinou com sucesso uma rede profunda 30 vezes maior do que até aí tinha sido reportado na literatura e tem, nos últimos dois anos, crescido em dimensão e alcance, com numerosos investigadores e grande diversidade de projectos. Este tipo de rede consegue simular (de forma estatística) o modo como nós captamos anomalias e fazemos julgamentos rápidos (Kurzweil, 2005).

Não tem, até agora, nenhum enfoque especial na investigação, simulação ou criação de CA, mas trabalha com diversas capacidades cognitivas e existe entre vários especialistas a opinião de que a Google é a empresa mais bem posicionada para (se isso for possível) vir a criar um organismo com consciência.

Numa perspectiva funcionalista, pode acontecer que as enormes redes neuronais do *Google Brain* venham a evoluir para dimensões equivalentes às de um cérebro humano[81] e que a consciência emirja da actividade coordenada de múltiplos mecanismos, o que, a acontecer, seria uma mudança de paradigma na compreensão do fenómeno e na resolução do problema duro. A *Google* adquiriu em 2014 a *DeepMind* cujas tecnologias incluíam, por exemplo, um modelo que mimetiza a memória de

[81] Neste momento a escala é a do cérebro de um rato, muito aquém dos 10^{12} neurónios e 10^{15} sinapses do cérebro humano.

curto prazo dos seres humanos e um "jogador" de Go *(AlphaGo*) que, em 2017, venceu o humano nº 1 mundial deste complexo jogo de estratégia. Todavia até ao momento não há qualquer sinal de consciência fenoménica nos diversos trabalhos dados à estampa, nem qualquer avanço na questão do problema duro.

6.8.2 Human Brain Project

O *Human Brain Project* (HBP), prodigamente financiado pela União Europeia, iniciado em 2013 a partir do projecto suíço "*Blue Brain",* tenta aplicar os princípios de engenharia inversa ao estudo do cérebro dos mamíferos, na tentativa de construir um modelo digital do cérebro, desde o interior da célula ao cérebro inteiro, passando pelas várias estruturas funcionais (Markram, 2006). Centra-se no nível neurobiológico e, segundo Segev (2013), entre outras grandes áreas, espera-se que possa ajudar a compreender a consciência, embora até à data nenhuma das publicações no âmbito do projecto refira implícita ou explicitamente a consciência. Conseguiu já reproduzir a anatomia celular, conectividade e comportamento eléctrico de uma pequena parte do neocórtex de um rato[82], mas sem incluir as células gliais, vasos sanguíneos, plasticidade, etc. Em 2015, sob pressão de centenas de cientistas, a Comissão Europeia recomendou a redução do alcance do projecto, limitando a investigação a temas realisticamente alcançáveis como por exemplo, o mapeamento pormenorizado do cérebro, deixando a simulação digital para segundo plano (Marquardt, 2015).

Para já, Markram (Gama, 2017) suscitou a hipótese de que nas redes neuronais do cérebro haja estruturas geométricas de até 11 dimensões de espaço, o que poderia ser uma das razões para a nossa dificuldade em o compreender

Um dos subprojectos é o *Episense* que investiga os mecanismos dos sentidos e da memória episódica, que é aquela que se relaciona com as nossas experiências conscientes no espaço e no tempo e que, no fundo, define o que somos.

Se o HBP, que deve concluir em 2023, conseguir atingir os ambiciosos objectivos iniciais (a reconstrução digital completa de um cérebro humano), é de prever que sejam dadas respostas conclusivas, num sentido ou noutro, sobre a o modo como a consciência é instanciada ou produzida pelo cérebro.

6.8.3 Iniciativa BRAIN

Lançada pela Administração norte-americana em 2013, a Iniciativa *BRAIN* (*Brain Research through Advancing Innovative Neurotechnologies),* tem também como objectivo compreender as funções do cérebro, mas só em 2016 arrancou a 1ª fase de desenvolvimento tecnológico e validação

[82] Um terço de um milímetro cúbico de tecido cortical.

(*Advisor Comittee to the Director*, 2013). Não há, até à data, qualquer publicação relevante na área da consciência.

6.8.4 Neurogrid

O *Neurogrid* (Boahen, s.d.) é um projecto da Universidade de Stanford especificamente desenhado para simular cérebros biológicos. Usa um misto de computação analógica e comunicação digital e consegue simular 1 milhão de neurónios e 6 biliões de sinapses, em tempo real. Não há até à data quaisquer publicações que, explícita ou implicitamente, abordem a consciência.

7 Discussão

Após a apresentação e sucinta análise dos modelos, teoria e projectos referidos, a ideia geral que retiro, é a de que se avançou muito na simulação de comportamentos e tarefas cognitivas associadas à consciência, mas não há nenhuma aproximação relevante à solução para o problema duro e decanta-se até uma razoável dúvida de que os aspectos subjectivos e qualitativos da experiência consciente possam sequer encaixar numa ontologia fisicalista. Na verdade nenhuma das implementações ou teorias que referimos, apesar das extraordinárias *performances* na simulação de comportamentos e na explicação de aspetos específicos associados à consciência, está hoje mais perto de instanciar a consciência, particularmente a fenoménica, do que estava o IBM 1401, lançado em 1962. Em todas as teorias e modelos afloram objecções várias à própria possibilidade de instanciação da CA. Objecções filosóficas, funcionais e experienciais que colocam mesmo em causa o paradigma ainda prevalecente nas ciências cognitivas e que se traduz na metáfora da mente e da consciência como sendo processamento de informação, e do cérebro como um computador.
Sobre o que ficou descrito nos capítulos anteriores, entendo especificamente que:

7.1 O paradigma materialista é dominante

Embora o idealismo não possa ser liminarmente refutado no quadro do actual paradigma científico, e o naturalismo não explique como é que as leis da física e da química "produzem" seres conscientes capazes de descobrir essas leis e perceber o universo que governam, a perspectiva materialista é ainda claramente dominante tanto na neurociência, como na IA/CA. Apesar de realmente não compreendermos os mecanismos pelos quais o cérebro sustenta as experiências subjectivas que estão no cerne de uma mente consciente, prevalece o otimismo quanto ao poder da ciência para os desvendar e quanto à factibilidade de máquinas verdadeiramente conscientes. A maioria dos cientistas acredita (Arrabales, 2016) que, leve o tempo que levar, e à semelhança do que sempre aconteceu com outros grandes enigmas com os quais o *sapiens* se confrontou, as respostas serão encontradas no estrito plano da matéria e das leis da física.

Contudo as correntes materialistas do fisicalismo e do funcionalismo enfrentam sérias objecções. A mais comum é a factual incapacidade para lidar com os aspectos qualitativos da experiência consciente já que as propriedades fenoménicas não fazem parte do mundo descrito pelas leis da física que conhecemos.

A analogia reducionista de Dennett (2017)[83] desemboca na absurda conclusão de que tudo aquilo que a cada um parece ser uma experiência consciente é apenas uma crença e só os objectos descritos

[83] O cérebro como uma máquina feita de biliões de máquinas (os neurónios)

pelas leis da Física serão reais. Absurda porque, uma vez que esse mundo "real" não é tudo o que experienciamos ou concebemos, forçoso é concluir que, ou nem tudo é Física, ou há uma Física diferente da que temos, como aliás sugere Roger Penrose, na sua *Orch-OR*, ou David Bohm, com a sua *Implicate Order*. Por outro lado, tanto estas como outras teorias quânticas e subquânticas acabam por ser abordagens que procuram preservar a cosmovisão materialista já que, embora não considerem a consciência um mero epifenómeno da actividade neuronal, procuram de facto circunscrevê-la ao mundo físico[84], mantendo assim em aberto a possibilidade de instanciar a CA.

De facto, uma possível física subjacente à *Orch-OR* e à *Implicate Order* pode eventualmente permitir o armazenamento, manutenção e manipulação da consciência usando *hardware* não orgânico, já que estamos em domínios abaixo do nível biológico e nestes uma partícula[85] é apenas uma partícula. Mas, sendo certo que não se pode objectivamente descartar a possibilidade de organismos artificiais conscientes, regidos por eventuais leis físicas, a verdade é que até hoje não se descobriu como e se as leis físicas que conhecemos se conectam com o fenómeno da consciência. Os mais pormenorizados mapeamentos dos mecanismos cerebrais até ao nível das moléculas, átomos ou *quarks*, não explicam como e porquê um processo físico dá origem a uma experiência subjectiva. Como vimos existe a esperança de que o HBP logre um completo mapeamento (neuronal) mas, até hoje, o que se pode honestamente dizer é que a observação de activação de um certo grupo de neurónios, associado à experiência subjectiva de um *quale*, não explica como é que esta experiência surge e por que razão é como é para aquele que a experiencia. O hiato explicativo entre os processos físicos biológicos e a experiência subjectiva não foi até agora superado o que, ressalve-se, não significa que não possamos instanciar a consciência trabalhando com processos físicos já que compreender algo não é, muitas vezes, condição necessária para realizar esse algo[86].

Quanto aos processos neuronais, se a mente fosse realizada apenas por processos neuronais no cérebro, como acredita a maioria dos neurocientistas, haveria que explicar por que razão há processos neuronais que são percebidos como mentais e outros não; Porque não há conteúdo mental num sonho sem sonhos, apesar de muitos processos neuronais estarem activos. A sugestão de que os estados neuronais idênticos a estados mentais poderão ter uma camada adicional de propriedades mentais, introduz algo que por sua vez necessita de outra explicação e desemboca afinal num dualismo de propriedades. Acresce o facto de que, se for válida a ideia funcionalista de que os processos mentais podem ser produzidos por diferentes substractos, como não podem deixar de acreditar os investigadores da IA que sustentam a possibilidade de instanciação da consciência em máquinas,

[84] Obviamente não apenas aquele que percebemos e captamos.
[85] Ou o que quer que exista.
[86] Por exemplo, certas leis físicas do mundo, como a teoria da gravidade de Newton, permitem cálculos muito precisos sobre o comportamento dos corpos, sem todavia nada adiantarem quanto à natureza da realidade. O mesmo se pode dizer da mecânica quântica, que funciona com leis muito precisas e úteis mas que sugerem uma realidade subjacente estranha e incompreensível. A um nível mais prosaico, os organismos biológicos "produzem" há milhões de anos, naturalmente, outros organismos biológicos, sem que compreendam os seus mecanismos.

teremos de concluir da irrelevância da sua natureza material o que, por tortuosos caminhos, nos reconduz a uma espécie de dualismo, segundo o qual o *hardware* cria o *software*, embora não se saiba como. De resto a ideia de que um conceito funcional captura a classe de todos os seus potenciais realizadores físicos e cada um destes tem umas propriedades físicas causais bem definidas, tais que todo o objecto que as tenha produz os mesmos efeitos, não é difícil de contestar. Pequenas variações (por exemplo a cor) em algumas das propriedades poderiam continuar a permitir que o realizador realize[87]. Poder-se-á tentar decantar eventuais propriedades relevantes, mas isso obriga a acrescentar uma cláusula impossível de definir em termos estritamente fisicalistas, o que anula, *ipso facto*, a tese da redutibilidade das propriedades funcionais às físicas.

7.2 O comportamento não basta para atribuir consciência

Quanto ao diagnóstico de consciência a partir de observação de comportamentos, é verdade que há quem entenda que, sendo o comportamento exterior a única indicação de estados fenoménicos[88], poderemos mesmo, segundo uma estrita lógica, ter de os atribuir a organismos que mimetizam o comportamento de organismo biológicos. Mas, embora alguns comportamentos animais pareçam efectivamente estar associados a estados fenoménicos, a verdade é que não temos dificuldade em conceber organismos que repliquem comportamentos conscientes sem que os experienciem. Vários *robots* avançados imitam comportamentos biológicos complexos e mesmo assim temos muita relutância em atribuir-lhes consciência fenoménica. Na inversa, qualquer bom actor consegue perfeitamente simular medo, dor, etc., sem experienciar esses *qualia*, e certos animais desenvolveram mesmo estratégias de sobrevivência, de si mesmos e das crias, que se baseiam na simulação comportamental. Ou seja, o comportamento por si só, quer seja por apreensão empática e holística, quer seja analisado por versões mais ou menos sofisticadas de Testes de Turing, não parece ser um indicador fiável da instanciação da consciência em máquinas. Como Searle e outros disseram, o TT, nas suas possíveis versões, não mede a consciência, mas apenas processamento de informação, particularmente a capacidade de seguir regras e imitar um certo estilo de comunicação. Visto isto, numa hipotética situação em que tanto um homem como uma máquina são igualmente inteligentes e interactivos, qual é a realmente marca da consciência? Na minha opinião não há qualquer marcador objectivo. Há apenas uma intuitiva, particular e indescritível constatação de algo inerente e

[87] Por exemplo, um afiador de facas pode variar em cor, tamanho, material de que é feito, etc., etc. Variar a cor não o impede de continuar a ser um afiador de facas.

[88] Harnad (2003) sublinha que é só através do comportamento que atribuímos consciência a outros sistemas e não através da observação da activação neuronal que, entende, é apenas um correlato do comportamento. Acreditamos que um organismo é consciente pelo que diz ou faz, e não pelo padrão de actividade do cérebro.

irredutível, que acontece quando encontro um ser senciente e reconheço isso com a minha consciência.

7.3 Teoria da Integração da Informação

A TII enquadra bem a unidade da consciência mas não explica como uma grande integração de informação produz as qualidades subjectivas com que cada um de nós experiencia o mundo. Os resultados actuais e conhecidos da TII evidenciam que a elevada integração de informação não basta para tornar um sistema consciente, o que é corroborado pelos casos de *split brain* que mostram que uma defeituosa integração da informação não impede os pacientes de experienciar *qualia* e ter absoluta consciência de si mesmos (Pinto, et al., 2017).

A TII, embora matematizada e susceptível de previsões e refutações, também não explica como é que a actividade neuronal dá origem a qualquer experiência subjectiva e por que razão o cérebro cria *qualia* se toda informação está já presente no circuito. Identificar as áreas do cérebro com mais elevado φ não é uma explicação, mas apenas a quantificação de uma correlação.

Maguire et al. (2014) argumentam que a integração computacional implica sempre perda de informação[89] e demonstram que a integração sem perdas requer funções não computáveis, o que leva a concluir que a consciência não pode ser computacional, nem modelada computacionalmente. Assim, se se aceitar que a consciência assenta na integração da informação os computadores não podem ser conscientes. *Ergo,* a consciência não seria possível nas máquinas. O problema duro persiste.

7.4 Correlatos Neuronais da Consciência

As teorias e modelos neuronais identificam eventuais CNC, mas nenhuma imagem da actividade cerebral mostra ou contém o modo como as coisas parecem ao organismo consciente, pese embora o prometedor trabalho de Nishimoto, et al.(2011)[90]. Em 2017 não é ainda clara a eventual contribuição específica para os conteúdos da consciência, de certos neurónios das áreas sensoriais primárias. Sabemos que, nos seres biológicos, o conteúdo específico de qualquer experiência consciente está correlacionado com actividade de certas partes do cérebro. Se essas partes forem lesadas, a pessoa pode perder a consciência de aspectos do mundo que elas processavam. Mas a imagem da actividade dos neurónios não é suficiente. Lesões graves no cerebelo não fazem desaparecer a consciência,

[89] A experiência do aroma do chocolate, num ser humano, é informação integrada com o resto das memórias e por isso um cirurgião terá dificuldade em erradicá-la, sem afectar todo o sistema neuronal. O mesmo não acontece com um detector artificial de cheiros, bastando cortar o acesso à base de dados.
[90] Os resultados destes autores são, de certa forma, circunscritos a uma circularidade subjectiva. As imagens que aparecem como "tradução" de padrões neuronais são apenas válidas para aquele sujeito e referidas à imagem real que o seu sistema perceptual capta. É, *mutatis mutandis*, uma versão do escaravelho de Wittgenstein. Não podemos saber se a neuro imagem é o resultado da actividade de um cérebro ou de uma mente que o examina.

apesar de haver aí mais neurónios que em qualquer outra parte do cérebro. Sabemos também que no começo de um sono profundo, a consciência se desvanece mesmo que os neurónios do sistema corticotalâmico mantenham a mesma actividade que no estado de vigília. Devido à intrincada conectividade do cérebro, a actividade no córtex prefrontal, nos gânglios basais e até no cerebelo, podem covariar sistematicamente sem que a percepção consciente seja a responsável. Na realidade nenhuma área do cérebro parece necessária para cada um de nós estar consciente (Koch, C. et al, 2016). De resto os CNC, por si mesmos, pouco informam relativamente à consciência de pacientes com severas lesões cerebrais, crianças, fetos, outras espécies, ou sistemas artificiais.

Parece pois que, no limite, é necessário algo mais do que meros correlatos. Se a consciência é de facto "produzida" pelo cérebro, alguns CNC têm de ser substractos essenciais da consciência e uma teoria completa tem de explicar os mecanismos, o porquê e o como dessas correlações. Até ao momento tudo isso é desconhecido, pelo que o problema duro permanece inacessível.

7.5 Outras teorias específicas

Relativamente aos automodelos, a ideia de que para haver experiência subjectiva é necessário um automodelo, ou uma forma especial de memória, implicaria que só cérebros complexos estariam aptos a experienciar subjectivamente. Olhando para a biosfera, a ideia parece-nos logo disparatada. E olhando para nós mesmos, há formas básicas de experiência subjectiva que nos aparecem como intrusões em processos mentais complexos. A dor súbita, por exemplo, é imperiosa, não pode ser ignorada e não se sente da maneira que se sente, por causa da complexidade cognitiva. Talvez se possa até experienciar dor sem um modelo interno do mundo, ou memória sofisticada, já que muitos animais com cérebros muito mais simples que o nosso, respondem à dor o que pode indicar que a sentem. Por outro lado, uma câmara de vídeo dirigida a um espelho, registando imagens de si mesma, não parece possuir consciência de si. Penso até que computadores e programas não podem ser verdadeiramente autorreferenciais, referem-se sempre a algo, e há sempre um conjunto de meta-regras que não são autorreferenciais, ao contrário da consciência.

É inegável que as teorias e implementações que incluem automodelos simulam bem certos aspectos da autoconsciência, mas não avançam um milímetro na que respeita ao problema duro. Como vimos, a própria teoria de Damásio limita-se a afirmar que os processos físicos que, na sua opinião, sustentam a consciência, levam à emergência desta através de *loops* autorreferenciais, mas não explica como isso acontece. Pode até dizer-se que é mais uma teoria do *self* do que da consciência, já que parece não abordar a experiência da experiência.

Quanto à emoção, o programa de Rosalind Picard assenta numa perspectiva cognitivista das emoções, modeladas como processamento de informação, reduzidas a sinais simples e discretos que

não captam a complexidade da experiência emocional. Já a provável complexidade resultante de abordagens que se centrem na interacção social, ambiental, cultural, contextual, etc., é muito resistente a formalização computacional. E, de resto, o caso de Phineas Gage (Damásio, 1994) parece provar que a consciência prevalece intacta, mesmo em situações em que certos mecanismos das emoções estão gravemente comprometidos.

Pelo seu lado as teorias de ordem superior neurologicamente motivadas mostram que se pode estudar e até simular a capacidade para reportar ou monitorizar as experiências próprias mas não se vê como as possam explicar.

A *Orch-OR* e a Teoria do Holofluxo sugerem que a solução do problema duro pode ter a ver com a própria natureza física da realidade, mas sublinham a necessidade de uma nova Física para que a hipótese possa ser comprovada. Penrose (1997) entende que se há algo na actividade física do cérebro que leva à consciência, tal não pode ser computacionalmente simulado e sugere mesmo que a não computabilidade pode ser uma característica da consciência.

Já a Teoria do Holofluxo assume implicitamente a ideia de que abaixo do nível quântico volta a haver determinismo e realidade objectiva, embora se possa colocar a questão sobre o que acontece a eventuais níveis ainda mais fundamentais, num ciclo interminável de níveis eternamente fora do nosso alcance cognitivo. Por outro lado, e de forma mais prosaica, se a percepção fosse holográfica, como a teoria sugere, não deveria haver casos em que a estimulação de células cerebrais produz falsas memórias (Ramirez et al., 2013), nem seria possível desenvolver maneiras de ver as estruturas que armazenam memórias em cérebros vivos (Gross et al., 2013).

Talvez seja ousado basear a mente e a consciência na mecânica quântica ou numa eventual física subquântica, teorias ainda não bem entendidas, e com muitos problemas e várias interpretações. Talvez seja até forçado concluir, como faz a Teoria do Holofluxo, que o cérebro opera segundo um modelo equivalente a um holograma, porque as pequenas ondas de Fourier, efectivamente encontradas nos estádios iniciais da percepção visual e auditiva, podem simplesmente ser uma natural decomposição deste particular tipo de sinal ambiental. Todavia, se Penrose estiver certo, não será através de algoritmos computacionais que chegaremos a máquinas conscientes, embora em teoria nada impeça a incorporação dos mesmos mecanismos físicos (até agora desconhecidos) eventualmente responsáveis pela nossa consciência, se existirem.

Já as teorias do ETG capturam bem certas observações cognitivas, mas definitivamente não mostram como é que uma determinada arquitectura pode levar à emergência de *qualia*.

7.6 A (não) computabilidade da consciência e o papel da intuição

Uma das objecções funcionais mais relevantes à instanciação da consciência em máquinas, releva do facto de os seres conscientes como nós, por exemplo, não operarem apenas sobre representacões simbólicas do mundo, ao contrário do que acontece com os computadores que efectivamente processam, armazenam e recuperam informação. Num computador uma imagem de uma paisagem fica realmente guardada, com todos os pormenores, num determinado módulo do sistema e pode ser transferida para outro sistema, sem quaisquer perdas. Tudo isto é guiado por algoritmos, ao contrário do que se passa connosco. O meu cérebro não armazena a canção que acabo de ouvir, completa, num qualquer ficheiro que se pode apagar, misturar, copiar ou transferir. O que o cérebro faz é a mudar a sua configuração física, embora pouco se saiba de como o faz[91]. A mesma canção pode fazer mudar de maneira diferente o cérebro de quem a ouve e nada há no processamento de símbolos que gere experiência subjectiva ou fenómenos psicológicos como as sensações qualitativas.

Todavia a metáfora computacional mantém-se poderosa e Kurzweill (2013) demonstra-o teorizando sobre "algoritmos no cérebro", vendo a mente inteligente como sendo o produto de uma máquina gigante, massivamente redundante, hierárquica, recursiva, que aprende por si mesma e faz predições baseadas na memória, refundando a visão de Minsky (1985) do cérebro como um "computador de carne". A perspectiva de Domingos, P. (2015), sobre um algoritmo universal, o *Master Algorithm* que deverá resultar, em matéria de *Machine Learning*, da grande unificação dos algoritmos de aprendizagem das abordagens conexionista, simbólica, evolutiva, analógica e *bayesiana*, ancora-se também nesta metáfora computacional.

A meu ver há aqui uma falácia lógica elementar, que consiste em concluir que todos os organismos capazes de comportamento inteligente são processadores de informação, a partir das premissas de que todos os computadores são capazes de comportamento inteligente e que todos os computadores são processadores de informação.

É verdade que muitos investigadores acreditam que os *robots* virão a ser conscientes (Arrabales, 2016) e que, num futuro em aberto, poderemos até transferir a "alma" para uma máquina, como Hofstadter (1981) faz literariamente com a mente de Einstein. Efectivamente uma vez a consciência reduzida a símbolos, poderia ser editada, copiada, vendida e até misturada e apagada. Esta é a ideia que subjaz, por exemplo, ao filme Transcendence (2014), no qual Johnny Depp representa o cientista cuja mente é transferida para a *internet*. Mas esta ficção, resultante de uma explicação funcional da mente e da consciência, suscita questões interessantes: quando ninguém consulta o *hardware* a consciência do sujeito está desactivada? Está entediado? A dormir? Se houver várias cópias, cada

[91] Num recente artigo de opinião Miller (2015) sugere que levará séculos apenas para ter uma ideia da conectividade neuronal básica, e milhões de anos para que seja possível fazer o *upload* ou recriar uma mente individual, a partir de um cérebro perfeitamente conservado.

uma delas experienciará a mesma coisa em cada momento? O mesmo "eu" experienciará a sua ubiquidade? Se uma cópia for destruída ou estiver desligada, desvanecer-se-á uma parte da consciência de si? Pode sentir dor? Medo? Paixão? Tudo ao mesmo tempo?

Não conseguimos responder a estas questões paradoxais porque simplesmente, até ao momento, ninguém sabe ao certo o que é a consciência, para além de algo inefável e enigmático que parece estar em cada um de nós e ser responsável pelo turbilhão dos pensamentos, pela sensação que brota quando vemos uma certa paisagem, ou ouvimos uma determinada música, ou perdemos um ente querido, etc.

Acompanho Dreyfus (1992), na ideia de que nem a inteligência humana é redutível a regras[92] nem o cérebro é uma versão complexa de um computador digital. As nossas capacidades parecem ir muito para além da computação. Não existe um algoritmo universal que decida se todas as Máquina de Turing (MT) se detém ou não, mas isso não nos limita a nós, possuidores de um cérebro.

Por outro lado, nenhum sistema formal finito pode completar um processo autorreferencial em tempo finito, como parece ser o caso da consciência. E todavia nós fazemo-lo constantemente, o que reforça a ideia de que se o cérebro é a sede da consciência, então, seja o que for, não é apenas uma MT. Gödel (1931) demonstrou que qualquer sistema formal de axiomas e regras de inferência, desde que seja suficientemente amplo para conter descrições de proposições aritméticas simples, e não tenha contradições, contém enunciados cuja verdade é indecidível, dentro do formalismo do sistema. Mas nós, usando intuições exteriores ao sistema, que estão na antítese do procedimento formal, somos capazes de decidir que esse especial enunciado é verdadeiro e, de um modo mais geral, alcançar enunciados verdadeiros, que não são deduzíveis pelos formalismos dos sistemas.

O argumento de Gödel torna patente a possibilidade de irmos para lá dos limites de qualquer sistema formal, pelo uso da intuição directa. Na verdade usamos a intuição do que é "evidente", até para decidir que axiomas ou regras usar para estabelecer um sistema formal, da mesma maneira que precisamos de intuições externas para decidir se este ou aquele algoritmo é adequado à resolução deste ou daquele problema. O matemático indiano Srinivasa Ramanujan, um caso típico dessa capacidade intuitiva, dizia que a "verdade" dos seus teoremas lhe surgia como se Deus os tivesse colocado na sua cabeça e não como resultado de um posterior processo formal de prova (Rajasekaran, 2014). Dirac (1982) garantia que foi o seu agudo sentido estético que lhe permitiu chegar à equação para o electrão[93] e a ciência está cheia de casos em que a intuição desempenhou um papel decisivo na descoberta e formulação de teorias[94].

Mas o que é a intuição?

[92] Porque depende de coisas como capacidades, corpo, emoções, imaginação e outros factores que não podem ser codificáveis em listas de factos.
[93] Equação complexa, aplicável a partículas como electrões e *quarks*, simétricas, com spin e massa, consistente com a mecânica quântica, com a teoria da relatividade especial, e que prevê a existência de antimatéria.
[94] Para um apanhado exaustivo de descobertas matemáticas alcançadas por intuição ou "inspiração", consultar Hadamard (1945)

Não parece que as intuições sejam elas mesmas algorítmicas ou sequer específicas da matemática (Gigerenzer, 2007). Os nossos juízos parecem assentar em combinações complexas e interligadas de dados sensoriais, raciocínios e conjecturas e, em muitas situações os critérios sobre o que é verdadeiro ou falso são eles mesmo variáveis e dificilmente encapsuláveis nos limites de qualquer sistema formal. BonJour (1998) pensa que a intuição racional é uma captação racional, um "ver" directo, imediato, não discursivo e não inferencial, de que uma proposição é necessariamente verdadeira. O facto de parecer necessariamente verdadeira constitui aliás a base da justificação *a priori*, que é uma fonte fundamental de justificação epistémica.

É verdade que Dawkins (1986), reducionista assumido, argumentou que o nosso "programa" pode ter evoluído por selecção natural ao longo de milhões de anos, mas não explica como é que este "programa" (a mente), que não parece ser, em si mesmo, um processo algorítmico, pode ajuizar sobre a validade de outros algoritmos. Além disso, se os nossos processos mentais conscientes consistissem apenas na activação de algoritmos, como defendem alguns proponentes da IA forte, as leis físicas que regulam o funcionamento do cérebro teriam pouca importância, porque qualquer dispositivo poderia executar o algoritmo. O facto é que o cérebro é um objecto físico, trabalha com leis físicas, usa electricidade e processa informação, tal como um computador, mas além disso é consciente. Numa perspectiva emergentista, pode ser apenas uma questão de complexidade (Theise & Kafatos, 2013), mas pode ter também a ver com a natureza da matéria e das leis físicas que regulam aquilo de que são constituídos os seres conscientes.

Como vimos, os mais prometedores projectos referidos no capítulo anterior assentam em RNA o que, de algum modo, reflecte a ideia de que o cérebro biológico é mais uma rede neuronal que se altera na interacção, do que um computador digital. Saliente-se todavia que mesmo as mais avançadas RNA são de uma grande simplicidade face às complexas características das redes neuronais biológicas, algumas das quais podem eventualmente vir a revelar-se essenciais para a eventual criação de modelos de consciência[95]. Para alguns a esperança reside na emergência de redes neuronais quânticas, superiores à computação clássica em certos problemas, pela redução da complexidade. Mas, sendo verdade que os computadores quânticos deverão ser mais rápidos em certos tipos de computação, não deixam de estar também limitados a operações algorítmicas e, como vimos, nem sequer é consensual, de um ponto de vista naturalista, que os processos quânticos sejam necessários para sustentar a consciência (Tegmark, 2014).

[95] Os modelos convencionais de RNA não incluem oscilações elétricas, sincronia *gamma*, conexões com células gliais, plasticidade neuronal, etc. Por outro lado, também não é claro que a consciência tenha a ver com essa complexidade, já que, por exemplo, o cerebelo, tem uma grande densidade de células neuronais e nada parece ter a ver com ela.

7.7 Hiato explicativo computacional.

Reggia et al. (2014) sugeriram que o insucesso na instanciação da CA se pode dever ao hiato explicativo, na sua versão computacional, simplificadamente a incapacidade de compreender como o processamento de alto nível da informação cognitiva[96] pode ser mapeado em computação neuronal de baixo nível, *i.e*, o tipo de computação que pode ser feito numa RNA. É uma questão puramente computacional, independente do *hardware* envolvido, seja ele artificial ou natural.

Num cérebro, não se pode "adivinhar" um pensamento, pela análise dos padrões de activação neuronal. Num computador, pelo *hardware* também não conseguimos "adivinhar" o *software*. A partir do exame da linguagem da máquina (basicamente "0s" e "1s"), não há como inferir com precisão o *software* que o computador está a executar, porque uma mesma sequência de "0s" e "1s" de um *software* pode ser interpretada de várias maneiras, consoante o que o *software* faz.

Na IA o hiato explicativo computacional relaciona-se também com o velho debate sobre os valores relativos das abordagens *top-down* (simbólico, numérico, etc.) e *bottom-up* (neuronal, enxame, etc.). Os métodos simbólicos têm êxito na modelação de tarefas cognitivas de alto nível, mas fracassam no reconhecimento de padrões, no controle motor de baixo nível, em contextos de incerteza, ruído na informação, etc. Já os métodos de neurocomputação *bottom-up* são eficazes e robustos na aprendizagem e classificação de padrões, controle de baixo nível, contextos com ruído na informação, incerteza, etc. Se fosse possível superar este hiato poderíamos comparar directamente os mecanismos neurocomputacionais associados a actividades cognitivas conscientes/relatáveis e ao processamento não consciente da informação, permitindo-nos determinar se há ou não correlatos computacionais da consciência, em analogia com os CNC (Cleeremans, 2005).

Tem havido alguns esforços no sentido de vencer este hiato[97], acreditando-se que a eventual identificação de convincentes correlatos neurocomputacionais poderia abrir uma via para a compreensão da natureza fundamental da consciência e, logo, a sua possível instanciação em organismos artificiais. Por outro lado, se nada se identificar, podem ser extraídas implicacões filosóficas que darão força às teorias que incorporam aspectos dualistas.

7.8 Objecto de estudo e instrumento de estudo

Um problema filosófico inultrapassável, pelo menos no que toca ao cérebro humano, é que no estudo dos fenómenos que dele emergem, o objecto de estudo coincide com o instrumento utilizado para o estudar. Não havendo uma contradição lógica, no que toca aos aspectos físicos, já em relação a certos fenómenos que dele supostamente emergem, o paradoxo é evidente. Para alguns o seu

[96] Algoritmos e estados dinâmicos usados para a resolução de problemas dirigidos a objetivos, tomada de decisões executivas, planeamento, linguagem e metacognição, etc., processos que se considera serem conscientes.
[97] Ver Reggia et al. (2014),

esclarecimento é apenas um limite prático actual (Churchland, 1995) mas para outros é definitivo, reside na própria natureza das coisas (McGinn, 1995)[98], quer seja porque os nossos limites cognitivos impossibilitam essa compreensão, quer seja porque a incapacidade é independente dos nossos ou outros limites cognitivos. Na verdade um sistema que se compreenda inteiramente a si mesmo, tem de se simular inteiramente a si mesmo e nenhuma simulação completa desse sistema pode estar contida nele. Seria o próprio sistema.

Mesmo algumas teorias quânticas que se invocam como esperança para alcançar a compreensão da natureza última dos processos mentais implicam, de acordo com as interpretações convencionais, que o cérebro tem de se observar a si mesmo, para que aconteça o colapso da onda descrita pela equação de Schroendinger.

De qualquer maneira a consciência fenoménica parece existir em organismos biológicos com cérebros mais simples que o nosso, pelo que, em teoria, pelo menos esses poderemos vir a estar aptos a estudar e descrever completamente em termos físicos e, numa perspectiva naturalista, compreender os mecanismos que sustentam o mundo fenoménico, se eventualmente existirem.

7.9 O Quarto Chinês, sintaxe e semântica

Proposta por Searle (1984), a experiência conceptual do "quarto chinês" consiste em imaginar uma pessoa num quarto isolado, recebendo perguntas em chinês, processando-as de acordo com regras escritas na sua língua natal e respondendo também em chinês, embora nada perceba de chinês. A ideia é que a actividade mental consiste simplesmente num sequência bem definida de operações (um algoritmo) que, por princípio, poderia funcionar em qualquer *hardware*. Os defensores da IA forte alegarão que, onde funcione, o algoritmo experienciará *qualia* e terá uma consciência. Será a mente. O que Searle diz é que apesar de este sistema exibir um comportamento associado a consciência e simular propriedades cognitivas atribuídas à consciência, não entende realmente, nem tem estados intencionais relativamente aos objectos representados nos caracteres chineses. Este argumento toma como premissas o facto de os programas de computador serem sintáticos, o facto de a sintaxe não ser condição suficiente para a semântica e o facto de as mentes terem semântica, concluindo que implementar um programa não basta para obter uma mente. Segundo esta linha de argumentação é completamente irrelevante a rapidez, a quantidade de memória, e a complexidade da programação. *Watson, Deep Blue, AlphaGo, Google Brain*, etc., serão apenas versões sofisticadas de uma Máquina de *Turing*, *i.e*, máquinas de manipulação de símbolos.

Efectivamente os computadores digitais são sistemas binários, o que significa que processam a informação em termos de dois estados, simbolizados por 1 e 0. E são estes dígitos que representam

[98] A corrente filosófica que defende esta posição designa-se *Misterianismo*

números, cores, formas, sons, etc. A operação de uma MT é sintática porque só reconhece símbolos, não o seu significado (semântica)[99]. Para Searle uma máquina que processa símbolos não é necessariamente uma máquina que compreenda símbolos.

Uma possível[100] resposta a este argumento contra a IA forte e contra o paradigma computacional da mente, é o "*symbol grounding*. Se os caracteres no quarto chinês puderem ser ligados a representações não simbólicas, como imagens ou sons, então o sistema como um todo pode compreender o significado dos símbolos e ter estados intencionais relativos a esse significado. Enquanto executa o algoritmo, pode começar a perceber algo da estrutura que formam os símbolos, sem compreender realmente o significado de muitos deles, individualmente. Por exemplo os caracteres chineses para "pão" poderiam ser substituídos por outro qualquer alimento, sem que a história fosse afectada de forma significativa. Ou seja, a implementação, em si mesma, terá um conteúdo semântico e pode ter um papel causal no mundo real, fazendo eventualmente emergir a consciência. Para Kuipers 2008), qualquer sistema que mantenha uma correspondência entre conceitos simbólicos de alto nível e um fluxo de dados de baixo nível, e que tenha um sistema de raciocínio que use estes símbolos alicerçados, tem experiências subjectivas verdadeiras, correspondentes a *qualia* e consciência de si.

Penso que nenhuma objecção ao argumento de Searle vence o hiato entre sintaxe e semântica e mesmo que o argumento não convença completamente, tal não muda o facto de que as MT são apenas manipuladoras de símbolos, sendo implausível que os significados reais das histórias sejam concretizados pela execução simples dos algoritmos.

Reconheça-se, todavia, que o argumento do quarto chinês não dá realmente nenhuma razão *a priori* para que o arranjo das partículas num computador seja menos capaz de consciência que o mesmo arranjo num cérebro. A já descrita experiência conceptual proposta por Chalmers (2010), sugere que a consciência não deverá desaparecer se cada componente de um cérebro humano fosse substituído por um equivalente artificial, gradual ou instantaneamente mas, como vimos, Tononi & Koch (2015), defendem exactamente o contrário.

7.10 Uma simulação é apenas uma simulação.

No âmbito da CA, o grande separador categorial é entre simulação e instanciação.

Quanto à simulação tem havido inegável sucesso na captura computacional de aspectos específicos da consciência, a partir dos seus correlatos neuronais, arquitecturais ou comportamentais. Mas, da mesma maneira que simular um furacão num computador, não é ter um verdadeiro furacão lá dentro,

[99] Saliente-se que até a palavra "reconhecer" é enganadora, porque implica uma experiência subjectiva.

[100] E algo esotérica

simular a consciência também não deve ser a consciência, porque uma simulação computacional de um fenómeno é apenas uma abstracção e simplificação, e não o próprio fenómeno. Modela-se o que se conhece, o que se pensa ser relevante, e o que é possível modelar, deixando de fora tudo o resto. A um nível prático também podemos ver a diferença entre uma experiência simulada dentro de uma simulação e a realidade que tenta simular e que existe fora da simulação. Se fizer uma simulação computacional de uma pessoa a comer um gelado, ela não prova realmente o chocolate, nem experiencia o que experienciam os seres reais que comem gelado. E mesmo que haja um grande número de níveis de simulação e que nós estejamos num deles, a consciência talvez seja o último nível de realidade, pelo que mesmo uma elevada simulação computacional da consciência não é a consciência e não convencerá um ser consciente de que está consciente. Isto é assim, a não ser que, no que toca a propriedades fenoménicas, de alguma forma a simulação fosse também replicação, mas esse seria um mero postulado dificilmente refutável.

Em suma, de alguma forma o *hardware* importa e as meras representações digitais não têm poder causal sobre nada no mundo.

Quando se fala de instanciação trata-se de ter realmente consciência fenoménica e experienciar *qualia,* mas até ao momento nada indica que se tenha avançado nesse sentido, e os *replicante*s de "*Blade Runner*", "*more humans than humans*", são apenas personagens de ficção. Talvez o desenho informado de um modelo com verdadeira CA (se for possível) exija um conhecimento preciso e completo da origem e da natureza da consciência humana. Conhecimento que por agora não existe, pelo que temos de nos contentar com vários modelos e teorias sectoriais, que misturam diferentes níveis de complexidade e conhecimento e não se podem confirmar ou refutar de forma clara.

O facto é que, até ao momento, a CA é fundamentalmente simulação e mesmo os *robots* com comportamentos complexos, em última análise têm a motivação de uma máquina de calcular. Face a isto, aqueles que acreditam na possibilidade de construir um organismo consciente, têm dois caminhos pela frente: ou copiar um que se saiba ser consciente, ou fazer evoluir uma máquina, acelerando o mecanismo evolutivo que se acredita ter desaguado no fenómeno da consciência.

Criar um cérebro, ou um modelo e um cérebro não tem dado, até agora, razão para optimismo: para citar apenas um exemplo o cérebro da lombriga *Caenorhabditis elegans* tem apenas 302 neurónios e já foram completamente mapeadas as suas 6000 sinapses e o diagrama das ligações (Ferris, 2012). Contudo, em 2017, não há nenhum modelo funcional deste exíguo sistema nervoso. Considerando as dimensões do cérebro humano, com centenas de triliões de sinapses em constante mudança pode-se ter uma ideia do que será modelar tal rede, na esperança de que, no fervilhar da complexidade, desponte a consciência.

Outro caminho é começar com arquitecturas abstractas, modelando o que se conhece ou se teoriza, relativamente aos mamíferos, e evolui-las até (espera-se) desembocar, por acaso e necessidade, num organismo consciente como se pensa ter, numa perspectiva evolucionista, acontecido connosco. Com isto em mente, temos de concluir que eventuais máquinas não biológicas conscientes terão de replicar os processos físicos, bioquímicos, processos analógicos celulares e moleculares, reacções químicas, forças electroestáticas, sincronias globais, *feedbacks*, *loops*, e conexões funcionais e estruturais que ocorrem no cérebro, e isto sem os materiais orgânicos que estão na base dos sistemas biológicos.

Embora nenhuma destas alternativas tenha, até hoje, produzido resultados encorajadores, faz todo o sentido que sejam exploradas e levadas até ao limite para que eventualmente se possa chegar, pela via empírica, a conclusões sobre a factibilidade de máquinas conscientes.

7.11 O problema da 1ª pessoa

Como ficou claro, estudar cientificamente a consciência encontra um limite no seu carácter irredutivelmente subjectivo. Não se consegue reduzir a vertente fenoménica a teorias sobre processamento de informação, sobre neurotransmissores, neurónios e estados do cérebro.

O estudo na perspectiva de outra pessoa, observando estados do cérebro, arquitecturas, comportamentos, relatos, etc., apenas permite fazer correlações, e não acede ao lado experiencial, por muito significativa que seja essa correlação. É verdade que temos hoje fortes medidas objectivas de estados qualitativos, na 3ª pessoa. Se alguém diz que está a sentir medo, pode-se verificar a resposta da amígdala com uma fMRI, medir a resposta galvânica nas mãos, o nível de cortisol no sangue, etc. Mas se muitas pessoas disserem que estão a sentir medo e não mostrarem nenhum destes sinais, as medidas "objectivas" deixarão de ser fiáveis porque o contravalor da mudança fisiológica é sempre o reporte consciente, na 1ª pessoa. E teremos sempre de confiar no relatório subjectivo das pessoas, para perceber se as correlações são precisas.

Mesmo as imagens como as de Nishimoto et al (2011), que, têm como pano de fundo a aspiração de resolver o problema mente-cérebro, identificando um padrão de disparo neuronal com uma certa imagem, dependem de que a pessoa reporte, em algum momento, o que está a ver ou a pensar e, no limite, apontam sempre para inferências estatísticas, além de que há muitos processos mentais diferentes que activam as mesmas áreas cerebrais, na mesma pessoa.

Não me parece pois possível abordar a consciência sem qualquer tipo de linguagem experiencial, interna, qualitativa, da mesma maneira que nenhum de nós pode abstrair da sensação de ser um "eu", que tem um corpo[101]. Há evidentemente quem entenda que a sensação de ser um "eu" e de haver um "*locus*" para ele, dentro do corpo, é apenas uma ilusão (Harris, 2014), um processo, não havendo

[101] Efectivamente a maioria das pessoas não sente ser um corpo, mas sim ter um corpo, mesmo que não seja capaz de descrever cognitivamente esse eu que tem um corpo

realmente qualquer "eu" unitário que seja transportado incólume de um momento para outro. Todavia sentimos ser esse "eu", que é o centro da experiência, que permanece ao longo da vida, mesmo quando todas as células mudaram, mesmo quando nenhum átomo é ainda o mesmo, mesmo quando, em certas situações, se parece situar fora do corpo, como acontece com milhões de pessoas em todas as culturas (Foe, 2007). Estas "ilusões", os sonhos lúcidos e outras do tipo místico, são também experiências, e têm forçosamente de dizer algo sobre a natureza da consciência. Algo que, pelo que desta discussão se depreende, parece não estar ao alcance do paradigma materialista uma vez que os próprios *qualia* artificiais são uma espécie de sucedâneo inteiramente dependente da perspectiva da 3ª pessoa, que é o observador humano, o qual apenas sabe ser consciente por observação na 1ª pessoa.

De resto nenhuma das abordagens da fenomenologia sintética até agora sugeridas especifica as qualidades ou modalidades sensoriais associadas aos conteúdos específicos da mente artificial nem até que ponto estes *qualia* poderiam ser análogos aos produzidos na experiência consciente de um ser humano.

8 Conclusões

É lícito concluir que não existe qualquer teoria ou definição consensual que explique e caracterize o fenómeno da consciência na sua totalidade. Muitas mentes se têm debruçado sobre o assunto, muitas teorias se elaboraram, muitos modelos se criaram e muitos mais se irão provavelmente criar mas, neste momento, não estamos em condições de dizer se se trata de um fenómeno intangível, de uma ilusão, de uma substância diferente da matéria, de uma produção da matéria, ou de qualquer outra coisa. Sabemos que, nos casos mais óbvios, parece correlacionada com o cérebro, mas pouco sabemos do como e do porquê. Pelo menos até ao momento, nenhum dos estudos examinados, ou quaisquer aproximações computacionais à CA apresentaram desenhos ou demonstrações convincentes da possibilidade de instanciação de consciência fenoménica num organismo artificial. Continuamos a não se saber como é que processos bioquímicos, eléctricos, físicos ou outros, criam sensações e experiência unificada.

Iniludível é o facto de que a consciência existe no mundo natural e cada um de nós a reconhece em si mesmo e suspeita nos outros. Seria uma contradição lógica dizer que não estou consciente de estar consciente, ou vice-versa, e parece-me evidente que não há ninguém que não experience a sua própria experiência. Por mera intuição estou bastante seguro de que outros humanos e animais são conscientes até um certo grau, e os *robots* sofisticados, o meu portátil ou a internet, não são. Não posso objectivamente provar a outrem que estou consciente ou que outros estão conscientes, mas tenho a convicção de que quem está a ler isto tem a íntima e invencível experiência de estar consciente. Se depende da matemática, da lógica, das leis da física, da química, da biologia, ou tem

origem em algo que transcende tudo isso, é, neste momento, uma pura questão de crença ou, se quisermos, de paradigma. Quem acredita no 1º caso não vê qualquer razão teórica para que a consciência não possa vir a ser reproduzida num organismo artificial, e crê que é uma questão de tempo.

Percebemos contudo que existem muitas objecções filosóficas, funcionais e experienciais a essa possibilidade. Algumas foram discutidas no capítulo anterior. Sendo objecções de monta, não desencorajam contudo a investigação, animada pela esperança de que, de alguma forma, a consciência fenoménica possa aparecer, nem que seja por acaso, numa implementação adequada. Não é, reconheça-se, uma esperança disparatada, já que os mecanismos do acaso e da necessidade estão profundamente arreigados na nossa cultura científica (Monod, 1971). Efectivamente, se a CA for possível e for uma propriedade emergente, mesmo que não seja fisicamente explicável, podem, por tentativa e erro ou por estratégias evolucionistas, ser criadas as condições que levem à sua emergência (Long & Kelley, 2007), o que seria decisivo na confirmação ou refutação da premissa emergentista.

Vimos que a maior parte do trabalho dos investigadores em CA se apoia em teorias consideradas essenciais no desenvolvimento de modelos computacionais da consciência. Demos aqui conta de novos métodos para raciocínio automatizado sobre automodelos, mudança de atenção com base em dados que não correspondem às expectativas, desenvolvimento do autorreconhecimento robótico, referimos a importância do "*symbol grounding*" como possível ponte entre computação neuronal de baixo nível e o raciocínio simbólico de nível superior, etc. Cada uma destas abordagens procura alicerçar-se em correspondentes correlatos neuronais, cognitivos, arquitecturais ou comportamentais da consciência, e cada uma delas representa uma posição teórica sobre a importância fundamental desses correlatos. Há que reconhecer que, do ponto de vista da CA simulada, tem havido progressos impressionantes, desde a criação de modelos computacionais que aumentam a activação do ETG, quando levam a cabo uma tarefa associada a esforços tidos como conscientes nos humanos, à inesperada identificação dos módulos de "*gating*" como os componentes mais "conscientes" de um neurocontrolo (Gamez D. , 2010), passando pela produção de notáveis comportamentos humanoides em robótica, como a ACH de Haikonen P. (2012) ou o *robot Sophia*, de David Hanson, pela demonstração de que organismos artificiais movidos pela expectativa, se podem reconhecer a si mesmos (Takeno, 2013), de que redes neuronais de segunda ordem podem mimetizar dados comportamentais de seres humanos com *blindsight* em certas tarefas (Pasquali et al., 2010), pela comprovação de que sinais de "*corollary discharge"* em modelos neurocomputacionais dos mecanismos *top down* de controlo da atenção podem dar conta do processamento consciente da informação (Taylor, 2012), etc.

Todavia também percebemos que no processamento de símbolos não parece haver nada que gere experiência subjectiva ou *qualia*. Tanto o cérebro como os computadores computam, mas só o cérebro parece compreender e experienciar, o que leva autores como Maguire et al. (2014), a sugerir que os *robots* nunca terão *qualia* e nunca desenvolverão consciência porque se esta faz qualquer coisa que a computação não faz, então a consciência fenoménica não poderá ser instanciada com processos computacionais. Não me parece que os obstáculos sejam apenas técnicos e subscrevo a tese de que a consciência não é processamento de informação e por isso poderá ser simulada por essa via, mas não instanciada.

Parece-me claro que a mente humana transcende a computação formal e usamos frequentemente intuições externas para decidir em tempo útil da verdade e validade de proposições, cálculos, procedimentos, etc. Filtramos quase instantaneamente o ruído da informação e manifestamente não somos meros sistemas formais. Agimos para lá da lógica das nossas crenças, dos nossos "programas", pensamos infinidades, fazemos coisas irracionais e espontâneas, imaginamos coisas que não estão nas premissas nem nos dados, algo que, por definição não estará jamais ao alcance do sonhado *Master Algorithm*[102].

Ficou também claro que a experiência subjectiva não parece requerer muitas coisas que tendemos a associar ao ser humano, como inteligência, emoções, memória[103], atenção[104], autorreflexão[105], linguagem[106], percepção e acção do mundo[107], pelo que a explicação final da consciência, se existir, não será uma explicação das capacidades cognitivas, mas sim dos fenómenos relacionados com a experiência subjectiva. E temos pois de assumir que também a CA, se for possível instanciá-la, poderá não necessitar de nada que um organismo biológico não necessite, incluindo as capacidades cognitivas referidas.

Vimos que, nos seres humanos, a falta de actividade neuronal é usualmente uma indicação de falta de consciência, mas o inverso não é verdadeiro, já que nem toda a actividade neuronal é percebida e relatada como experiência subjectiva[108], pelo que a detecção de actividade neuronal por terceiros não serve como prova de consciência. O mesmo deve acontecer com eventuais organismos artificiais cuja actividade interna "cerebral" poderá ser inspeccionada com instrumentos, sem que daí se possa concluir se há (ou não) experiência subjectiva. Na verdade pode não ser possível determinar

[102] Segundo Domingos o *Master Algorithm* será capaz de aprender tudo a partir dos massivos dados disponíveis.
[103] O caso de estudo do paciente HM mostra que se pode perder completamente a memória e mesmo assim ter experiências conscientes.
[104] Conduzir um carro num circuito habitual, por exemplo.
[105] Quando estamos absorvidos num *trekking* numa montanha, estamos vividamente conscientes do mundo, sem qualquer necessidade de reflexão ou introspecção e a imagiologia prova que podemos estar conscientes mesmo quando as áreas do córtex correlacionadas com a representação de si mesmo estão parcamente activas.
[106] Além dos outros animais, há muitos pacientes incapazes de usar ou perceber palavras e contudo estão conscientes e podem reportar as suas experiências de outro modo.
[107] Quando sonhamos estamos desconectados do meio ambiente, mas estamos conscientes. Stephen Hawking está consciente, apesar de completamente imobilizado pela doença de Gherig.
[108] Quando estou a dormir tenho actividade neuronal mesmo que não esteja a sonhar e a actividade do meu cerebelo não é consciente.

objectivamente, sem margem para qualquer dúvida, se um organismo artificial é ou não consciente, o que nos força a aceitar a possibilidade teórica de consciência num organismo artificial, ainda que não o possamos saber. Mesmo que um computador passe uma sofisticada versão do TT e pareça tão inteligente e consciente como um ser humano, poderá não estar realmente ciente de o ter passado, nem ter qualquer experiência da sua própria realidade ou ser. Poderá ser apenas uma máquina complexa e inanimada[109].

No limite as hipóteses são bastante simples: A instanciação da consciência em organismos artificiais é possível ou não, e tanto a produção filosófica como o trabalho nos campos da Neurociência, da IA e da CA, nos encaminham, *à la longue*, para esta disjunção exclusiva.

Quanto à singularidade e aos cenários apocalípticos, se algum dia viermos a descobrir o que permite a um objecto físico tornar-se consciente, é concebível que possamos construir organismos com CA. Estes organismos teriam sobre nós a vantagem de se poderem desenhar especificamente para serem conscientes, sem terem de crescer a partir de uma célula, sem o lastro das partes "inúteis" da evolução. Mas não estamos perto disso. Os melhores organismos artificiais que até hoje conseguimos modelar têm a consciência fenoménica de uma máquina de calcular.

Por outro lado, também não foram apresentadas evidências científicas indiscutíveis sobre a impossibilidade de instanciar a consciência em máquinas, embora no âmbito de algumas teorias, como a TII, se venham produzindo bons argumentos nesse sentido.

No lato campo das possibilidades tudo está em aberto, sendo igualmente possível que a consciência não seja uma criação do cérebro, resultante da selecção natural, por mutação aleatória de ADN. É possível que não seja sequer matéria/energia mas um processo organizador, complexificador, a montante da evolução biológica. É também possível encará-la de uma forma quase teísta[110], consistente com a mente universal do budismo, com o "Inconsciente Coletivo" de Jung, com o dualismo cristão, etc. Pode tratar-se de algo que releva de um propósito teleológico que não nos é acessível, de uma manifestação do princípio antrópico[111] ou pura e simplesmente de um fenómeno material cuja explicação exige a descoberta de novas leis físicas. São perspectivas que, por definição, não estão ao alcance do método científico tal como hoje o entendemos mas, no actual estado da arte, são tão possíveis como as que nele se baseiam.

Quanto a mim considero que há uma certa *hubris* em acreditar que podemos fabricar a consciência com alguns gramas de silicone, electricidade e regras.

[109] Literalmente, "sem alma".

[110] Embora as outras também sejam compatíveis com a ideia de Deus.

[111] O princípio antrópico defende que a natureza do Universo em que existimos está fortemente condicionada pela exigência de que devem estar presentes seres como nós, para a observar.

Mesmo sem entrar nas complexidades matemáticas das estranhas teorias de Bohm (1990) e Pribram (2013) sobre a consciência como podendo pertencer a uma *Implicate Order*, que subjaz a uma *Explicate Order*, na qual se situariam as nossas percepções da realidade, do tempo e do espaço, acredito que a consciência nem é um epifenómeno, nem emerge da realidade física. Acredito que faz parte da fábrica da realidade e é irredutível a qualquer outra coisa. Tal como a energia, conceito que não se pode coisificar, não detectamos directamente, nem conseguimos definir em função de outras "coisas", a consciência parece ser tão fundamental, tão axiomática, que é impossível de provar, tocar ou descrever em função de outra coisa, embora cada ser senciente pareça ter uma experiência directa e íntima da sua. Mas se consciência for de facto um constituinte básico e fundamental da realidade, então jamais seremos capazes de explicar como emerge de itens não conscientes, porque simplesmente não emerge.

Por tudo isto acredito que está fora do alcance um organismo artificial que, além das clássicas capacidades cognitivas, bom senso e julgamento moral, tenha livre arbítrio, jogue por antecipação, preveja os futuros. Que pense em sentido lato, que faça a gestão do risco em situações perigosas. Que tenha consciência do que está a fazer (Coelho, 2008), e que, além de tudo isto, experiencie subjectivamente o mundo.

É sintomático que conclua com palavras como "acredito" ou "parece", etc. Para além do que no capítulo 7 escrevi sobre a intuição como fonte de justificação epistémica, Gigerenzer (2007) dizia que a intuição não é acerca de saber a resposta certa de repente, mas sim de entender instintivamente qual a informação relevante e irrelevante, no contexto de um assunto que se estudou profundamente. Tenho para mim que intuir não é adivinhação mágica nem uma aposta de lotaria, mas sim o conjunto de juízos que fazemos continuamente enquanto estamos conscientes, reunindo factos, impressões, conhecimentos, memórias, emoções, etc., num processo que não parece ser algorítmico (ou, caso o seja, não é executável em tempo útil).

Ao escrever este trabalho, nada do que li e analisei contém provas racionais concludentes sobre a possibilidade ou impossibilidade da CA, no sentido fenoménico, pelo que me resta a intuição. E a minha intuição é que nem o cérebro é uma mera Máquina de Turing, nem a consciência é uma propriedade emergente da matéria, ou do modo como esta se organiza, mas sim algo de mais fundamental a montante, imanente e irredutível. Parece-me também que não estamos mais perto de instanciar a consciência fenoménica em máquinas do que estávamos em 1968, quando o HAL 9000 era um artefacto da ficção científica, inspirado por Marvin Minsky e é esta a minha resposta à questão suscitada no título.

9 Referências

Adami, C. (2006). What do robots dream of? *Science, 314,* 1093-1094.

Advisor Committee to the Director. (2013). *Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group.* National Institutes of Health.

Aleksander, I. (2009). Essential Phenomenology for Conscious Machines: A Note on Franklin, Baars and Ramamurthy: "A Phenomenally Conscious Robot". *APA Newsletter on Philosophy and Computers, 8* (2).

Aleksander, I., & Dunmall, B. (2003). Axioms and tests for the presence of minimal consciousness in agents. *Journal of Consciousness Studies, 10,* 7-18.

Aleksander, I., & Gamez, D. (2009). Iconic training and effective information. *Proceedings of the AAAI fall symposium BICA II* (pp. 2-10). AAAI.

Aleksander, I., & Gamez, D. (2011). Informational theories of consciousness: A review and extension. In C. Hernandez, R. Sanz, J. Gomez, L. Smith, A. Hussain, A. Chella, & I. Aleksander (Eds.), *From brains to systems: Brain-inspired cognitive system*. Berlin: Springer.

Alexander, E. (2012). *Proof of Heaven*. Simon & Schuster Paperbacks, New York.

Arrabales, R. M. (2011). *Evaluation and development of consciousness in artificial cognitive systems (Doctoral thesis)*. Madrid, Spain: Universidad Carlos III.

Arrabales, R.M (2016). *Most people think robots will become conscious.* Obtido em 10 de Maio de 2017, de Conscious-Robots.com: http://www.conscious-robots.com/tag/poll

Arrabales, R., Ledezma, A., & Sanchis, A. (2010). The cognitive development of machine consciousness implementations. *International Journal of Machine Consciousness, 2*, 213-235.

Ascoli, G. (2005). Brain and mind at the crossroads of time. *Cortex*, 619-620.

Baars, B. (1997). In the Theatre of Consciousness: Global Workspace Theory. A Rigorous Scientific Theory of Consciousness. *Journal of Consciousness Studies*, 4, 292-309.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Science, 6*, 47-52.

Baars, B., & Franklin, S. (2007). An architectural model of conscious and unconscious brain function. *Neural Networks, 20*, 955-961.

Baars, B., Ramsov, T., & Laureys, S. (2003). Brain, conscious experience and the observing self. *Trends in neurosciences, 26 nº 12*, 671-675.

Balduzzi, D., & Tononi, G. (2009). *Qualia*: the geometry of integrated information. *PLoS Comput. Biol, 5*. Obtido de http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000462

Barron, A., & Klein, C. (2016). What insects can tell us about the origins of consciousness. In M. Gazzaniga (Ed.), *Proc Nati Acad SCI USA*, (pp. 4900-4908).

Berkeley, G. (1988). *Principles of Human Knowledge and Three Dialogues between Hylas and Philonous*. London: Penguin.

Block, N. (1978). Troubles With Functionalism. *Midwest Studies in the Philosophy of Science,* 9: 261–325.

Block, N. (1995). On a confusion about the function of consciousness. *Behavioural and Brain Sciences, 18*, 227-47.

Block, N. (2002). Concepts of Consciousness. In D. Chalmers, *Philosophy of Mind- Classical and contemporary readings* (pp. 206-218). New York: Oxford University Press.

Block, N. (2007). Consciousness, Accessibility and the mesh between psychology and neuroscience. *Behavioural and Brain Sciences, 30*, 481-548.

Boahen, K. (s.d.). *Brains in Silicon.* Obtido em 09 de Maio de 2017, de http://web.stanford.edu/group/brainsinsilicon/index.html

Bohm, D. (1980). *Wholeness and the Implicate Order*. London: Routledge.

Bohm, D. (1990). A New Theory of the Relationship of Mind and Matter. *Philosophical Psychology 3* (2): 271–86.

Bohm, D., Basil J. H. (1993). *The Undivided Universe: An Ontological Interpretation of Quantum Theory*. London: Routledge.

Boltuc, P. (2009). The Philosophical Issue in Machine Consciousness. The Behavioural and Brain Sciences, 18(2), 227-287.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science, 314*, 1118-1121.

BonJour, L. (1998). *In Defense of Pure Reason*. Cambridge, MA: Cambridge University Press.

Bosse, T., Jonker, C., & Treur, J. (2008). Formalization of Damasio's theory of emotion, feeling and core consciousness. *Consciousness and Cognition, 17*, 94-113.

Buschman, T., & Miller, E. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science, 315*, 1860-1862.

Carruthers, P. (2000). *Phenomenal Consciousness: A naturalistic theory.* Cambridge: Cambridge University Press.

Casali, A. G. et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med*. 5, 198ra105

Cattel, R., & Parker, A. (2012). Challenges for brain emulation. *Natural intelligence, 1*, 17-31.

Chalmers, D. (1995). Facing up the problem of consciousness. *Journal of Consciousness Studies, 2*, 200-19.

Chalmers, D. (1996). *The Conscious Mind.* Oxford: Oxford University Press.

Chalmers, D. (2003). The content and epistemology of phenomenal belief. In A. Jokic, & Q. Smith (Eds.), *Consciousness: New Philosophical Perspetives*. Oxford: Oxford University Press.

Chalmers, D. (2010). The singularity: A Philosophical Analysis. *Journal of Consciousness Studies, 17*, 7-65.

Chalmers, D. (2017). Singularity and dualism. *Entrevistado por Adriana Graça,* em 07 de Fevereiro de 2017.

Chella, A., & Gaglio, S. (2009). In Search of Computational Correlates of Artificial *Qualia*. *The 2nd Conference on Artificial General Intelligence*.

Chella, A., & Gaglio, S. (2012). Synthetic phenomenology and high-dimensional buffer hypothesis. International *Journal of Machine Consciousness, 4,* 353-365.

Chella, A., & Macaluso, I. (2006). Sensations and perceptions in "Cicerobot" a museum guide robot. *Proceedings of BICS.* Lesbos.

Chella, A., Frixione, M., & Gaglio, S. (2008). A cognitive architecture for robot self- consciousness. *Artificial Intelligence in Medicine, 44,* 147-154.

Chrisley, R. (2009). Synthetic Phenomenology. *International Journal of Machine Consciousness, 1 nº 1*, 53-70.

Chrisley, R., & Parthermore, J. (2007). *Robotic Specification of the Non Conceptual Content of Visual Experience*. Sussex: COGS / Department of Informatics, University of Sussex Falmer.

Churchland, P. S. (1995). *The Engine of Reason and Seat of the Soul*. Cambdrige, MA: MIT Press.

Ciompi, L. (2003). Reflections on the role of emotions in consciousness and subjectivity, from the perspetive of affect-logic. *Consciousness & Emotion, 4, nº 2*, 181-196.

Cleeremans, A. (2005). Computational correlates of consciousness. In S. Laureys (Ed.), *Progress in brain research,* (Vol. 150, pp. 81-98).

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and metarepresentation: a computational sketch. *Neural Networks, 4*, 1032-1039.

Clowes, R., & Seth, A. (2008). Axioms, properties and criteria: roles for synthesis in the science of consciousness. *Artificial Intelligence in Medicine, 44,* 91-104.

Coelho, H. (2008). *Teoria da Agência: Arquitectura e Cenografia*, Lisboa

Cotterhill, R. (2003). CyberChild. A Simulation Test-Bed for Consciousness Studies. *Journal of Consciousness Studies, 10 nº 4*-5, 31-45.

Crick, F. & Koch, C. 1990. Toward a Neurobiological Theory of Consciousness. *Seminars in Neuroscience, 2*: 263–75.

Crick, F. (1994). *The Astonishing Hypothesis. The Scientific Search for the Soul*. New York: Scribners.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature neuroscience, 6,* 119-126.

Crick, F., & Koch, C. (2005). What is the function of the claustrum? Philos. Trans. R. Soc. Lond. B Biol. Sci, 360, 1271-1279.

Damásio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. NY: Avon.

Damásio, A. (1999). *The feeling of what happens.* NY: Harcourt Brace & Company.

Damásio, A. (2003). *Ao Encontro de Espinosa: As emoções sociais e a neurologia do sentir.* Círculo de Leitores.

Damásio, A. (2011). *Self Comes to Mind: Constructing the Conscious Brain*, Random House

Dawkins, R. (1976). *The selfish gene.* Oxford University Press.

Dawkins, R. (1986). *The blind watchmaker.* London: Longman.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Ng, A. (2012). Large Scale Distributed Deep Networks. *Neural Information Processing Systems.*

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the USA, 95*, pp. 14529-14534.

Dehaene, S., Sergent, C., & Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings National Academy of Sciences, 100,* pp. 8520-8525.

Dennett, D. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Penguin Books.

Descartes, R. (1637). *The Discourse on the Method.*

Dirac,P. (1982), Pretty mathematics, *Int. J. Theor. Phys.*, 21 pp. 603-605.

Doan, T. (2009b). *Conscious-Robot.com.* Obtido de http://www.conscious-robots.com/2009/12/10/pentti-haikonens-architecture-for-conscious-machines/

Domingos, P. (2015).The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, New York

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, Massachusetts: MIT Press.

Edelman, G. (1989). *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books

Hampson R., Song D., Opris I., Santos L., Shin D., Gerhardt G., Marmarelis V., Berger T., Deadwyler S. (2013). Facilitation of memory encoding in primate hippocampus by a neuroprosthesis that promotes task-specific neural firing. *Neural Eng. 2013; 10(6).*

Ferris, J. (2012). *The connectome Debate: Is Mapping the Mind of a Worm Worth It?* Obtido em Maio 10, 2017, de Scientific American: https://www.scientificamerican.com/article/c-elegans-connectome/

Fischer, D., Boes, A., Demertzi, A., et al. (2016). A human brain network derived from coma-causing brainstem lesions. *Neurology,* 87:23, 2427-2434.

Fodor, J. (1974). Special Sciences. *Synthese, 28*, 77-115.

Fodor, J. (1985). Précis of "The Modularity of Mind." *Behavioral and Brain Sciences, 8,* pp. 1-42.

Foe, A. (2007). *Consciousness Beyond the Body: Evidence and Reflection.* Kindle Edition.

Franklin, S., Strain, S., Snaider, J., McCall, R., & Faghini, U. (2012). Global workspace theory, its LIDA model, and the underlying neuroscience. *Biologically Inspired Cognitive Architectures, 1*, 32-43.

Gama, J. (2017). El cerebro funciona hasta en once dimensiones. *ABC.* Obtido em 12 de Junho de 2017 de http://www.abc.es/ciencia/abci-cerebro-funciona-hasta-once-dimensiones-201706121448_noticia.html

Gamez, D. (2005). An ordinal probability scale for synthetic phenomenology. In R. Chrisley, R. Clowes, & S. Torrance (Ed.), *Proceedings of the AISB05 symposium on next generation approaches to machine consciousness,* (pp. 85-94). Hatfield, UK.

Gamez, D. (2006). The XML approach to synthetic phenomenology. In R. Chrisley, R. Clowes, & S. Torrance (Eds.), *Proceedings of the AISB05 symposium on next generation approaches to machine consciousness* (pp. 85-94). Hatfield, UK.

Gamez, D. (2007). *What We Can Never Know: Blindspots in Philosophy and Science.* London: Continuum International Publishing Group.

Gamez, D. (2008, September). Progress in machine consciousness. *Consciousness and cognition, 17 nº 3*, 887-910.

Gamez, D. (2010). Information integration based predictions about the conscious states of a spiking neural network. *Consciousness and Cognition, 19*, 249-310.

Gross, G.G., J.A. Junge, R.J. Mora, H.B. Kwon, C.A. Olson, T.T. Takahashi, E.R. Liman, G.C. Ellis-Davies, A.W. McGee, B.L. Sabatini, et al. (2013). Recombinant probes for visualizing endogenous synaptic proteins in living neurons. *Neuron.* 78:971–985. doi:10.1016/j.neuron.2013.04.017

Gazzaniga, M. (1988). *Mind Matters: How Mind and Brain Interact to Create our Conscious Lives.* Boston: Houghton Mifflin.

Gennaro, R. (2012). *The Consciousness Paradox.* Cambridge, MA: MIT Press.

Gigerenzer, G. (2007). *Gut Feelings, The Intelligence of the Unconscious*. Penguin Books.

Godel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter System I, in *Monatshefte fur Mathematik una Physik, 38*, pp. 173-198 (Tradução de B. Meltzer, B., Dover Publications INC. New York).

Goertzel, B. (2011). Hyperset models of self, will and refletive consciousness. *International Journal of Machine Consciousness, 3,* 19-53.

Hadamard, J. (1945). *The psychology of invention in the mathematical field.* Princeton University Press.

Haikonen, P. (2007a). *Robot brains: circuits and systems for conscious machines*. John Willey & Sons.

Haikonen, P. (2009). The role of associative processing in cognitive computing. *Cognitive Computation, 1*(1), 42-49.

Haikonen, P. (2012). *Consciousness and robot sentience*. World Scientific.

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the Orch OR theory. *Physics of Life Reviews, 11,* 39-112.

Hameroff, Stuart R., Travis J., Tuszynski, J. (2014). Quantum Effects in the Understanding of Consciousness. *Journal of Integrative Neuroscience (13)* 2: 229–52. doi:10.1142/S0219635214400093.

Harnad, S. (1992). The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *SIGART Bulletin*, 3(4), 9-10.

Harnad, S. (2003). Can a machine be conscious) How? in O. Holland, *Machine Consciousness.* Exeter: Imprint Academic.

Harnad, S., & Scherzer, P. (2007). First, Scale Up to the Robotic Turing Test; Then Worry About Feeling. *AI and Consciouness: Theoretical Foundations and Current Approaches. FS-07-01*, pp. 72-77. AAAI Fall Symposium Technical Report.

Harris, S. (2014). *Waking Up: A Guide to Spirituality Without Religion* . Simon & Schuster.

Hart, J., & Scassellati, B. (2011). Robotic models of self. In M. Cox, & A. Raja (Eds.), *Metareasoning* (pp. 283-293). MIT Press.

Hasker, W. (1999). *The Emergent Self.* NY: Cornell University Press.

Hawking, S. (2017). Stephen Hawking Issues Stern Warning On AI: Could Be 'Worst Thing' For Humanity, *Forbes,* consultada em 07 de Novembro de 2017, em https://www.forbes.com/sites/johnkoetsier/2017/11/06/stephen-hawking-issues-stern-warning-on-ai-could-be-worst-thing-for-humanity/#784fd51553a7

Hoel, E. (2017). *Agent Above, Atom Below, How agents causally emerge from their underlying microphysics*. New York: Department of Biological Sciences, Columbia University.

Hofstadter, D. (1981). A Conversation With Einstein's Brain. In D. Hoftsadter, & D. Dennett (Eds.), *The Mind's I: Fantasies and Reflections on Self and Soul.* Bantam Books.

Holland, O. (2007). A Strongly Embodied Approach to Machine Consciousness. *Journal of Consciousness Studies, 14*, 97-110.

Holland, O., Knight, R., & Newcombe, R. (2007). A robot-based approach to machine consciousness. In A. Chella, & R. Manzotti (Eds.), *Artificial Consciousness*. Exeter: Imprint Academic.

Husserl, E. (1960). *Cartesian Meditations.* Translated by Dorion Cairns. The Hague: Nijhoff.

Jackson, F. (1982). Epiphenomenal *qualia*. *Philosophical Quarterly, 32,* 127-136.

Jordan, J. (1988). *Synthetic phenomenology? Perhaps, but not via information processing*. Max Planck Institute for Psychological Research, Munich, German, Munich, Germany.

Joye, SR. (2016). *The Pribram-Bohm Holoflux Theory of Consciousness,* San Francisco, CA

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8 (*5), 679-685.

Key, B., Arlinghaus, R., & Browman, H. (2016). Insects cannot tell us anything about subjective experience or the origin of consciousness. *Proc Natl Acad Sci USA, 113 (27),* p. E3813.

Kinsbourne, M. (1988). Integrated field theory of consciousness. In A. Marcel and E. Bisiach, eds.*Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Kitamura, T., Tahara, T., & Asami, K. (2000). How can a robot have consciousness? *Advanced Robotics, 14*, 263-275.

Koch, C., & Tononi, G. (2011). A Test for consciousness. *Scientific American*, 44-47.

Koch, C., Massimini,M., Boly, M., Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience* 17, 307–321 (2016)

Koubeissi, M., Bartolomei, F., Beltagy, A., & Picard, F. (2014). Electrical stimulation of a small brain area reversibly disrupts consciousness. *Epilepsy & Behavior, 37*, 32-35.

Kuipers, B. (2008). Drinking from the firehouse of experience. *Artificial Intelligence in medicine, 44*, 155-170.

Kurzweil, R. (2005). *The Singularity is near.* New York: Viking Books.

Kurzweill, R. (2013). *How to Create a Mind: The Secret of Human Thought Revealed.* New York: Penguin Books.

Lamme, V. 2006. Toward a true neural stance on consciousness. *Trends in Cognitive Science 10:11*, 494–501.

Levine, J. (2014). Modality, semantics, and consciousness. *Philos Stud, 167*, 775-784.

Lindsay, O., Alexander, M., & Derek, M. (2015). Blindsight and subjective awareness of fearful faces: Inversion reverses the deficits in fear perception associated with core psychopathic traits. *Cognition & Emotion, 29* (7), 1256-1277.

Ling, G., & Spencer, L. (2017). Development of Grammatical Accuracy in English-Speaking Children With Cochlear Implants: A Longitudinal Study. *Journal of Speech, Language & Hearing Research*, 1062-1075.

Llinas, R. (2001). *I of the vortex: from neurons to self.* Cambridge, MA: MIT Press

Long, L., & Kelley, T. (2007). Review of Consciousness and the Possibility of Conscious Robots. *Journal of Aerospace Computing, information and communication, 7.*

Lycan, W. (1996). *Consciousness and Experience.* Cambridge, MA: MIT Press.

Lycan, W. (2009). Higher-order representation theories of consciousness. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 346-350). Oxford University Press.

Maguire, P., Moser, P., Maguire, R., & Griffith, V. (2014). Is consciousness computable? Quantifying integrated information using algorithmic information theory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Ed.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin: Cognitive Science Society.

Manzotti, R. (2012). The computational stance is unfit for consciousness. *International Journal of Machine Consciousness, 4*, 401-420.

Markram, H. (2006). The Blue Brain Project. *Nat. Rev.Neuroscience, 7*, 153-160.

Marquardt, W. (2015). Human Brain Project Mediation Report (2015), Juelich, Germany.

Massimi, M., Ferrarelli, F., Huber, R., Esser, S., Singh, H., & Tononi, G. (2005). Breakdown of Cortical Effective Connectivity During Sleep. Science, 309(5744), pp. 2228-2232.

Mc Dermott, D. (2007). Artificial intelligence and consciousness. In P. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 117-150). Cambridge University Press.

McGinn, C. (1995). Consciousness and space. In T. Metzinger, *Conscious Experience.* Paderborn: Ferdinand Schoningh.

Metzinger, T. (2000a). The subjectivity of subjective experience. In T. Metzinger (Ed.), *Neural correlates of consciousness* (pp. 285-306). MIT Press.

Miller, K. (2015, Oct 10). *Will You Ever Be Able to Upload Your Brain?* New York Times.

Minsky, M. (1985). *The Society of Mind*. New York: Simon & Schuster, Inc.

Minsky, M. (2006). *The Emotion Machine.* Simon & Schuster. ISBN 0-7432-7663-9.

Monod, J. (1971). *O acaso e a necessidade*. Publicações Europa-América.

Monti, M., Vanhaudenhuyse, A., Coleman, M., Boly, M., Pickard, J., Tshibanda, L., Laureys, S. (2010). Willful Modulation of Brain Ativity in Disorders of Consciousness. *The New England Journal of Medicine.*

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435-436. Retrieved Jan 2017

Nagel, T. (1979). Panpsychism. In T. Nagel, *Mortal Questions.* Cambridge: Cambridge University Press.

Nagel, T. (2012). Mind and Cosmos. Why the Materialist Neo-Darwinian Conception of Nature is Almosr Certainly False: Oxford University Press

Neckar, M., & Bob, P. (2016). Synesthetic associations and psychosensory symptoms of temporal epilepsy. *Neuropsychiatric Disease and Treatment. NIH, 12*, 109-112.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. (2011). *Reconstructing Visual Experiences from Brain Ativity Evoked by Natural Movies*. University of California, Berkeley. Berkeley, CA: Current Biology. doi: DOI 10.1016/j.cub.2011.08.031

Cook,N., Damásio, A., Carvalho,G.,Hunt, H., & Sacks, O. (2015). *Carta Aberta a Christopher Koch*, obtida em 10 de Outubro de 2017 em https://blogs.scientificamerican.com/mind-guest-blog/exclusive-oliver-sacks-antonio-damasio-and-others-debate-christof-koch-on-the-nature-of-consciousness/

Oh, H., Gentili, R., Reggia, J., & Vidal, J. (2012). Modeling of visuospatial perspetives processing and modulation of fronto-parietal network ativity during action imitation. *Proceedings 34th annual international conference of the IEEE engineering in medicine and biology society.*

O'Reagan, J. (2007). How to Build Consciousness into a Robot: The Sensorimotor Approach. *50 years of Artificial Intelligence*, 332-346.

Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition, 117*, 182-190.

Penrose, R. (1997). *The Large, the Small and the Human Mind.* New York: Cambridge University Press.

Persaud, N., Mcleod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 17, 257-261.

Picard, R. (1997). *Affective computing.* MIT Press, Cambridge (1997)

Pinto, D., Neville, D., Otten, M., Corballis, P., Lamme, V., Haan, E., Fabri, M. (2017). Split brain: divided perception but undivided consciousness. *Brain.* doi:10.1093/brain/aww358

Pizzi, R., Giuliano, S.,Fiorentini, S.,Pappalardo, V.,Pregnolato, M. (2010). Evidences of New Biophysical Properties of Microtubules. In *Artificial Neural Networks,* edited by Seoyun J. Kwon. New York: Nova Science Publishers, 1–17. Retrieved from https://air.unimi.it/retrieve/handle/2434/167480/168890/evidences.pdf.

Posner, J., Saper, C., Schiff, N., & Plum, F. (2007). *Plum and Posner's diagnosis of stupor and coma*. Oxford University Press.

Pribram, K. (2013). *The Form Within: My Point of View*. Westport, CT: Prospecta Press.

Prinz, J. (2003). Level-headed mysterianism and artificial intelligence. In O. Holland (Ed.), *Machine Consciousness*. Exeter: Imprint Academic.

Projects in Affective Computing (2017). Obtido em 25 de Outubro de 2017 em http://affect.media.mit.edu/projects.php, Massachusetts Institute of Technology,

Putnam, H. (1967). The nature of mental states. In Capitan, & Merrill (Eds.), *Art, Mind and Religion.* Pittsburgh: University of Pittsburgh Press.

Pylkkänen, P. (2007). *Mind, Matter, and the Implicate Order*. New York: Springer

Raffone, A., & Pantani, M. (2010). A global workspace model for phenomenal and access consciousness. *Consciousness and Cognition, 19*, 580-596.

Rajasekaran, G. (Director). (2014). *Ramanujan* [Motion Picture].

Ramirez, S., X. Liu, P.-A. Lin, J. Suh, M. Pignatelli, R. L. Redondo, T. J. Ryan, S. Tonegawa (2013). Creating a False Memory in the Hippocampus. *Science, 2013;* 341 (6144): 387 DOI: 10.1126/science.1239073

Rangarajan, V. et al.(2014). Electrical stimulation of the left and right human fusiform gyrus causes different effects in conscious face perception. *J. Neurosci*. 34, 12828–12836.

Reggia, J. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks, 44*, 112-131.

Reggia, J., Monner, D., & Sylvester, J. (2014). The computational explanatory gap. *Journal of Consciousness Studies, 21 (9)*, 153-178.

Rosenthal, D. (2005). *Consciousness and Mind.* Oxford: Oxford University Press.

Sacks, O. (2012). *Hallucinations.* Picador

Samsonovich, A., & Ascoli, G. (2005). The conscious self: ontology, epistemology and the mirror quest. *Cortex, 41*, 621-636.

Samsonovich, A., & Dejong, K. (2005). A general-purpose computational model of the conscious mind. In M. Lovett, & et al. (Ed.), *Proceedings of the sixth international conference on cognitive modeling* (pp. 382-383). ICCM-2004.

Seager, W. (2004). A Cold look at HOT Theory. In R. Gennaro, *Higher-Order Theories of Consciousness*. Philadelphia: John Benjamins.

Searle, J. (1984). Mind, brains and programs. *Behavioural and Brain Sciences, 3*, 417-457.

Searle, J. (1992). *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.

Searle, J. (1997). *The Mystery of Consciousness* (Vol. 44 (4)). New York: The New York Review of Books.

Searle, J. (2016, July 15). *Where Does Consciousness Come From?* TED Radio Hour, USA.

Segev, I. (2013). ASC 2012: Prof. Idan Segev - The blue brain". The Hebrew University of Jerusalem. Obtido de https://en.wikipedia.org/wiki/Blue_Brain_Project#cite_note-ASC_2012-3:

Seth, A. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition, 17*, 981-983.

Seth, A. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*, 1, 71-82.

Seth, A., Baars, B., & Edelman, D. (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition, 14 nº 1*, 119-139.

Seth, A., Izhikevich, E., Reeke, G., & Edelman, G. (2006). Theories and measures of consciousness. *Proceedings of the National Academy of Sciences, 103,* pp. 10799-10804.

Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition, 15*, 433-449.

Shanahan, M. (2010). *Embodiment and the Inner Life.* Oxford University Press.

Shanahan, M. (2015). *Ascribing Consciousness to Artificial Intelligence.* London: Department of Computing, Imperial College.

Singer, W. (1999).Neuronal synchrony: a versatile code for the definition of relations. *Neuron, 24*: 49–65.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies, 10*, 133-172.

Sokolov, I., Naumova, N., Nees, J., & Mourou, G. (2010). Pair Creation in QED-Strong Pulsed Laser Fields Interacting with Electron Beam. *Physical Review Letters, 105* (19).

Sporns, Q. (2011). *Networks of the brain.* MIT Press.

Stening, J., Jacobsson, H., & Ziemke, T. (2005). Imagination and Abstraction of Sensorimotor Flow: Towards a Robot Model. In R. Chrisley, S. Clowes, & S. Torrance (Ed.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness.* Hatfield.

Strawson, G. (1994). *Mental Reality*. Cambridge, MA: MIT Press, Bradford Books.

Sylvester, J., Reggia, J., Weems, S., & Bunting, M. (2013). Controlling working memory with learned instructions. *Neural Networks, 41*, 23-38.

Takeno, J. (2008). A robot succeeds in 100% mirror image cognition. *International Journal on Smart Sensing and Intelligent Systems, 1*, 891-911.

Takeno, J. (2013). *Creation of a conscious robot.* Pan Stanford.

Taylor, J. (2003a). Neural models of consciousness. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 263-335). MIT Press.

Taylor, J. (2007). CODAM: a neural network model of consciousness. *Neural Networks, 20*, 983-992.

Taylor, J. (2012). Does the corollary discharge of attention exist? *Consciousness and Cognition, 21*, 325-339.

Taylor, J., & Fragopanagos, N. (2007). Resolving some confusions over attention and consciousness. *Neural Networks, 20*, 993-1003.

Tegmark, M. (2014). *Consciousness as a State of Matter.* Chaos, Solitons & Fractals.

Theise, N.,Kafatos, M. (2013). Sentience Everywhere: Complexity Theory, Panpsychism & the Role of Sentience in Self-Organization of the Universe, *Journal of Consciousness Exploration & Research*, Volume 4, Issue 4,pp. 378-390

Thomson, H. (2014). Woman of 24 found to have no cerebellum in her brain. *New Scientist*.Obtido em 03 de Setembro de 2017 em https://www.newscientist.com/article/mg22329861-900-

Tinsley, C. (2008). Using topographic networks to build a representation of consciousness. *BioSystems, 92*, 29-41.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*, 42.

Tononi, G. (2008). Consciousness as Integrated Information: a Provisional Manifesto. *The Biological bulletin, 3*, 216-242.

Tononi, G. (2012). The integrated information theory of consciousness: an updated account. *Arch. Ital. Biol.* 150, 56–90.

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience, 4, 31*.

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B, 370.*

Treisman, A. (2009). Attention: theoretical and psychological perspetives. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 180-204). MIT Press.

Van Gulick, R. (2004). Higher-order global states HOGS: an alternative higher-order model of consciousness. In R. Gennaro (Ed.), *Higher-Order Theories of Consciousness.* Philadelphia: John Benjamins.

Van Gulick, R. (2016). *Consciousness.* (E. Zalta, Ed.) Retrieved April 14, 2017, from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/win2016/entries/consciousness

Van Lommel, P., van Wees, R., Meyers, V., & Elfferich, I. (2001). Near-death experience in survivors of cardiac arrest: a prospetive study in the Netherlands. *The Lancet, 358*, nº 9298, 2039-2045.

Vinge, V. (1993). *On the singularity.* (S. D. University, Ed.) San Diego: Department of Mathematical Sciences.

Wallace, R. (2006). Pitfalls in biological computing: canonical and idiosyncratic dysfunction of conscious machines. *Mind and Matter, 4, 91*-113.

Ward, L. (2011). The thalamic dynamic core theory of conscious experience. *Consciousness and Cognition, 20*, 464-486.